Cerner-VMware Customer Support Day
Kansas City, MO

Rupinder Saini, Sr. Manager Global Support Services (GSS)

May 25, 2011

vmware

Agenda

10:30AM	**Welcome/Kick-Off** Brian Stuckey, Director of Infrastructure Services Operations, Cerner
10:45 AM	**GSS Overview** *Rupinder Saini, GSS Senior Manager, VMware*
11:15 AM	**vStorage Best Practices** Paul Clark, Sr. Escalation Engineer, VMware
12:15 PM	**Lunch - Q&A with Experts**
12:45 PM	**Migration to ESXi** Ben Thomas, Sr. Federal Technical Support Engineer, VMware
1:45 PM	**Break**
2:00 PM	**Performance Troubleshooting and Best Practices** Ben Thomas, Sr. Federal Technical Support Engineer , VMware
3:15 PM	**Q&A with Experts**
3:45 PM	***Closing Remarks and Giveaways***

Welcome/Kickoff

Brian Stuckey, Director of Infrastructure Services Operations, Cerner

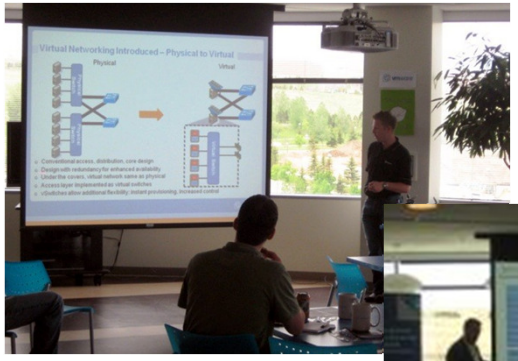
GSS Overview

Rupinder Saini, Sr. Manager Global Support Services

VMware Customer Support Days

Bringing VMware Support Experts, Sales & Customers Together

- **Customers learn directly from our experts -- GSS Senior Tech Support Engineers**
- **Learning event -- sharing of practical Best Practices, Tips and Tricks, and Top SRs/Issues**
- **In 2010, VMware held 24 Customer Support Days involving more than 800 customers globally**
- **Held quarterly across the world at Support Centers and on the Road**
- **Topics are driven by customer input & feedback**



http://www.vmware.com/support/customer_days.html

VMware Global Support Services: Mission and Value Proposition

Be trusted by customers and partners to ensure their success by delivering industry-leading, world-class services; be a competitive asset for the company.

World's Largest Virtualization Support Organization

Nearly 650 support staff in 6 Support Centers…
1,000s including partners

Supporting 250,000 Customers

2010- Handled 1.4 Million calls and 415K Service Requests

Twelve Years of Experience

- Supporting complex, production and development environments
- Supporting heterogeneous (Windows & Linux) environments

Global Support Services

Global Coverage 24x7, 365 days/year

Follow-the-sun Support for Severity 1 Issues

Support Relationships with 100% of the Fortune 100; 98% of Fortune 500

Global Support Services Goal

100% Customer Satisfaction

- Fast response times
- Aggressive resolution times
- Provide access to technical information
- Deliver enterprise-class support offerings
- Provide global focus
- Be an easy company to work with
- Be a company that listens
  - CSAT survey reviews and customer feedback

249,000 Global Support Requests in 2009
155,000 Americas Support Requests in 2009
8.9 (out of 10) CSAT, Americas

Questions

vStorage: Troubleshooting Performance

Paul Clark, Sr. Escalation Engineer, VMware

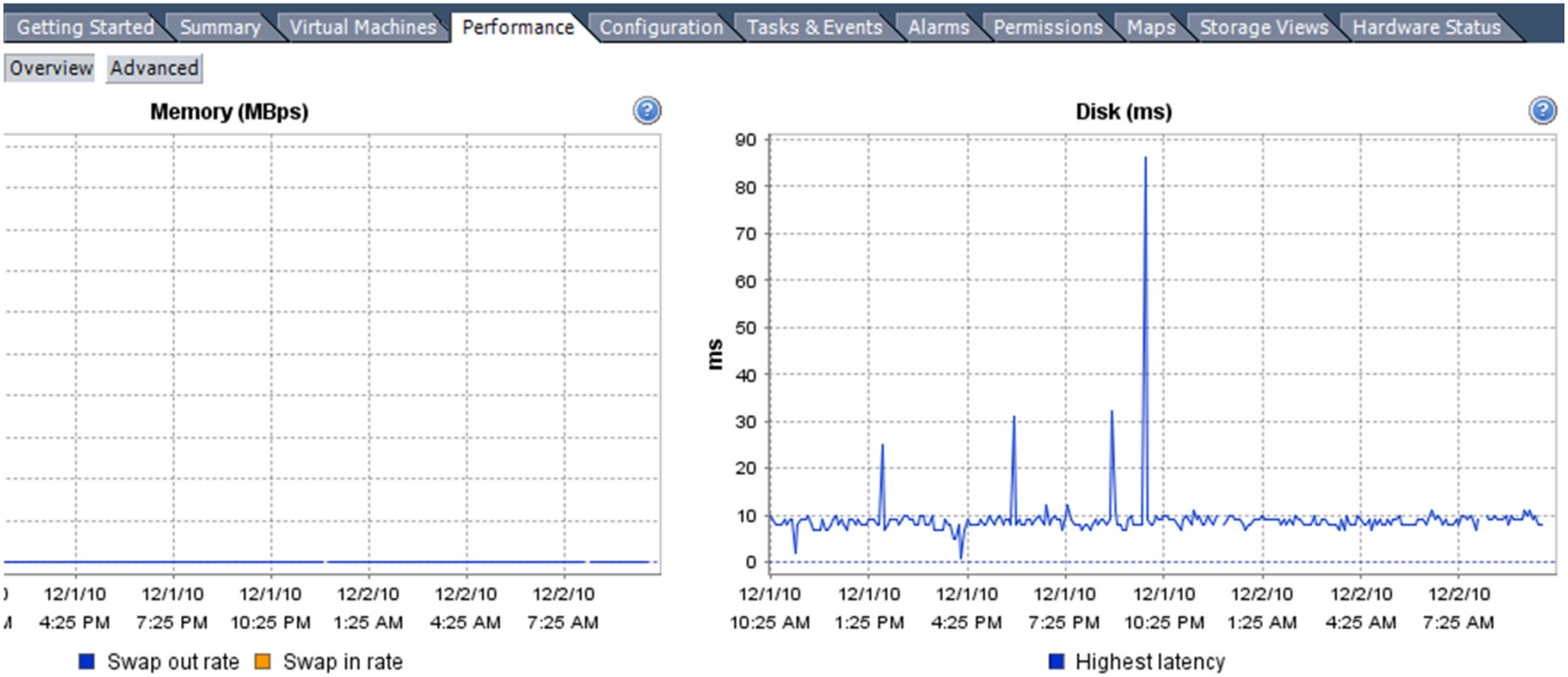
Agenda

- vCenter Performance Charts, ESXTop
- SCSI Reservations
- Multipathing Considerations

vCenter Performance Charts & ESXTop

Performance Charts

- Through the Overview section of the Performance tab of an ESX host, we can observe if a host is seeing recent disk latency issues.



- Unfortunately, this chart does not pinpoint which device or path is seeing this latency. For a more granularity, we will use “Advanced”

Performance Charts

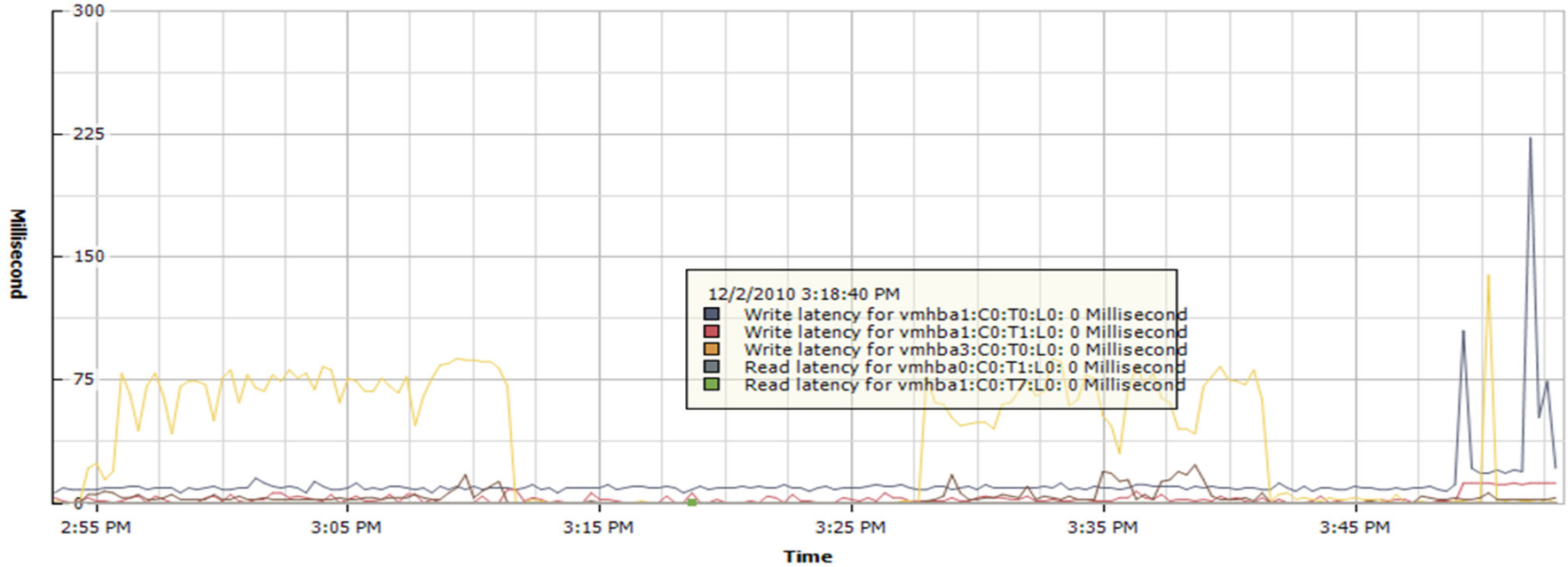
- If we switch to the "Storage path" option, we can see latency is coming from vmhba0:C0:T0:L0 for both read and write latency:

Overview | Advanced

Storage path/Real-time, 12/2/2010 2:53:15 PM - 12/2/2010 3:53:15 PM Chart Options...

Switch to: Storage path

Graph refreshes every 20 seconds

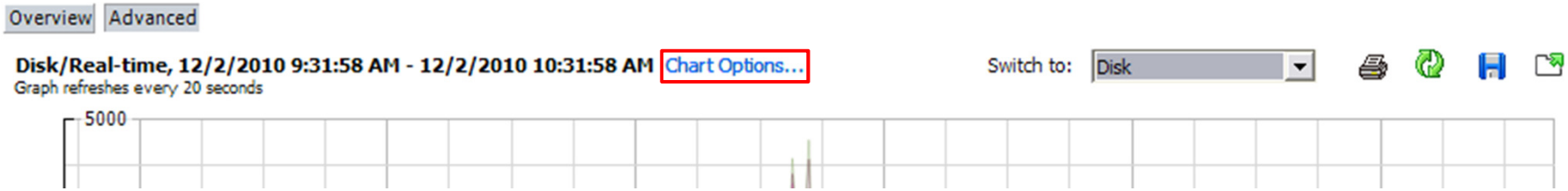


Performance Chart Legend

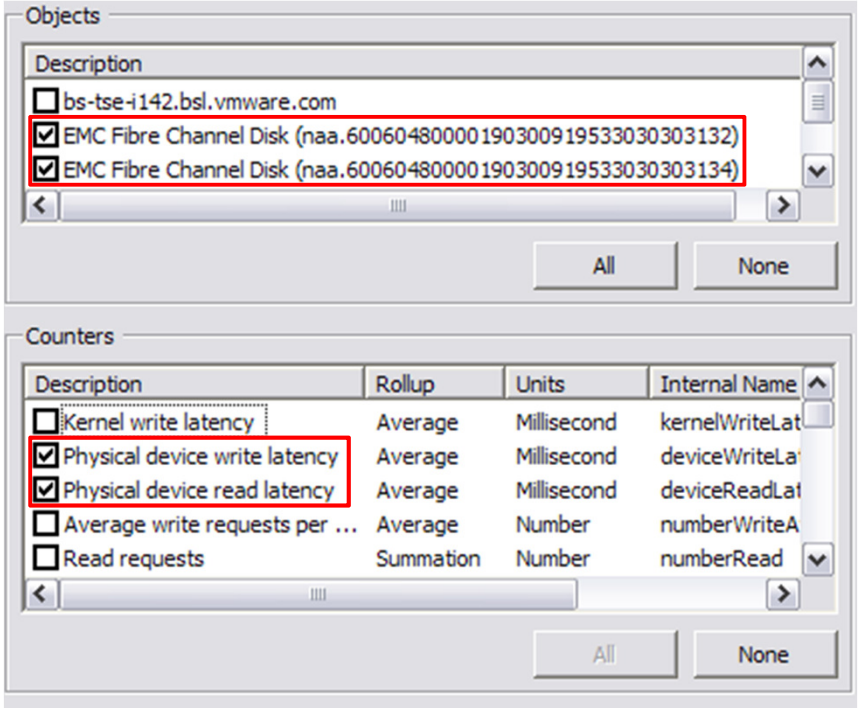
Key	Object	Measurement	Rollup	Units	Latest	Maximum	Minimum	Average
	vmhba1:C0:T5:L0	Read latency	Average	Millise...	0	0	0	0
	vmhba1:C0:T2:L0	Write latency	Average	Millise...	0	0	0	0
	vmhba0:C0:T0:L0	Write latency	Average	Millise...	21	223	6	11.839
	vmhba0:C0:T0:L0	Read latency	Average	Millise...	12	12	0	2.417
	vmhba1:C0:T3:L0	Read latency	Average	Millise...	0	0	0	0
	vmhba1:C0:T7:L0	Write latency	Average	Millise...	0	0	0	0

Performance Charts

- In order to see the latency statistics for a given LUN, we will need to go to the "Disk" view and then select "Chart Options":



Select the LUNs you wish to see the latency statistics for under "Objects", then select the Physical device read and write latency "Counters"



Performance Charts

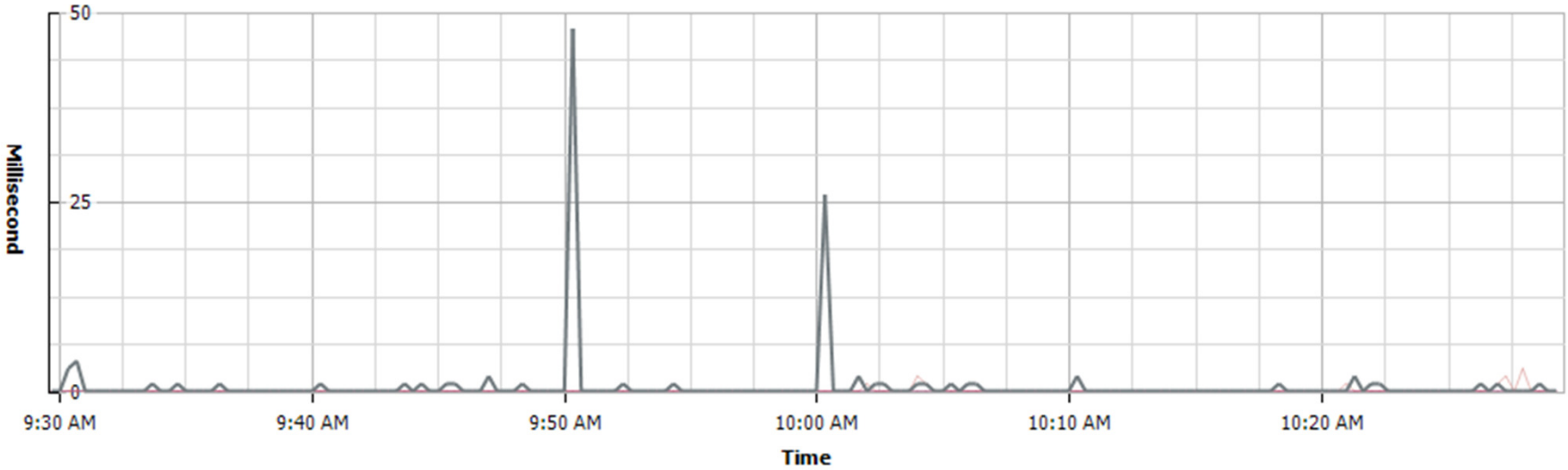
- With this granular view, we can see that the LUN with ID naa.60060480000190300919533030303134 had a latency spike up to 48ms at ~9:50am

Overview | Advanced

Disk/Real-time, 12/2/2010 9:29:34 AM - 12/2/2010 10:29:34 AM Chart Options...

Switch to: Default

Graph refreshes every 20 seconds



Performance Chart Legend

Key	Object	Measurement	Rollup	Units	Latest	Maximum	Minimum	Average
	EMC Fibre Channel Disk (naa.60060480000190300919533030303132)	Physical devicewrite...	Average	Millise...	0	0	0	0
	EMC Fibre Channel Disk (naa.60060480000190300919533030303134)	Physical devicewrite...	Average	Millise...	0	3	0	0.061
	EMC Fibre Channel Disk (naa.60060480000190300919533030303132)	Physical deviceread...	Average	Millise...	0	0	0	0
	EMC Fibre Channel Disk (naa.60060480000190300919533030303134)	Physical deviceread...	Average	Millise...	0	48	0	0.628

ESXTop

- ESXTop is a great tool to use if you need performance captured during a specific timeframe (scripted batch mode) or if the ESX host is unavailable in vCenter/VI Client.
- To view the equivalent of the “Storage path” view, hit ‘d’:

```
10:38:46am up 31 days 22:16, 228 worlds; CPU load average: 0.24, 0.23, 0.25

 ADAPTR PATH                    NPTH   CMDS/s  READS/s WRITES/s MBREAD/s MBWRTN/s DAVG/cmd KAVG/cmd GAVG/cmd QAVG/cmd
 vmhba0 -                          2    80.91     0.00    80.91     0.00     1.43     9.60     0.01     9.61     0.00
 vmhba1 -                          9    22.49     0.60    21.88     0.01     0.25     0.84     0.01     0.85     0.00
 vmhba2 -                          0     0.00     0.00     0.00     0.00     0.00     0.00     0.00     0.00     0.00
 vmhba3 -                          1     0.60     0.00     0.00     0.00     0.00     1.65     0.02     1.68     0.00
vmhba32 -                          0     0.00     0.00     0.00     0.00     0.00     0.00     0.00     0.00     0.00
vmhba33 -                          0     0.00     0.00     0.00     0.00     0.00     0.00     0.00     0.00     0.00
```

ESXTop

- You can toggle the display fields by hitting ‘f’. By taking out unneeded fields (I/O stats), you can make the display screen far easier to read and work with:

```
Current Field order: ABCdefGhijkl

* A:  ADAPTR = Adapter Name
* B:  PATH = Path Name
* C:  NPATHS = Num Paths
  D:  QSTATS = Queue Stats
  E:  IOSTATS = I/O Stats
  F:  RESVSTATS = Reserve Stats
* G:  LATSTATS/cmd = Overall Latency Stats (ms)
  H:  LATSTATS/rd = Read Latency Stats (ms)
  I:  LATSTATS/wr = Write Latency Stats (ms)
  J:  ERRSTATS/s = Error Stats
  K:  PAESTATS/s = PAE Stats
  L:  SPLTSTATS/s = SPLIT Stats

Toggle fields with a-l, any other key to return:
```

```
11:03:02am up 31 days 22:40, 228 worlds; CPU load average: 0.40, 0.34, 0.27

 ADAPTR PATH                    NPTH DAVG/cmd KAVG/cmd GAVG/cmd QAVG/cmd
 vmhba0 -                          2     8.65     0.01     8.66     0.00
 vmhba1 -                          9     0.42     0.02     0.44     0.00
 vmhba2 -                          0     0.00     0.00     0.00     0.00
 vmhba3 -                          1     1.74     0.03     1.77     0.00
vmhba32 -                          0     0.00     0.00     0.00     0.00
vmhba33 -                          0     0.00     0.00     0.00     0.00
```

ESXTop

- To view individual LUN statistics (with unique ID), hit ‘u’:

```
11:07:14am up 31 days 22:44, 228 worlds; CPU load average: 0.30, 0.33, 0.29

DEVICE              PATH/WORLD/PARTITION DQLEN WQLEN ACTV QUED %USD  LOAD   CMDS/s  READS/s WRITES/s MBREAD/s MBWRTN/s
mpx.vmhba3:C0:T            -                1     -    0    0    0  0.00     3.81     0.00     0.00     0.00     0.00
naa.5000c5000bf            -               64     -    0    0    0  0.00     0.00     0.00     0.00     0.00     0.00
naa.5000c5000bf            -               64     -    0    0    0  0.00     0.00     0.00     0.00     0.00     0.00
naa.60060480000            -               32     -    0    0    0  0.00     0.00     0.00     0.00     0.00     0.00
naa.60060480000            -               32     -    0    0    0  0.00     0.00     0.00     0.00     0.00     0.00
{NFS}Burlington            -                -     -    0    -    -     -     0.00     0.00     0.00     0.00     0.00
{NFS}vmlibrary             -                -     -    0    -    -     -     0.00     0.00     0.00     0.00     0.00
```

- As we can see, the default field size for “DEVICE” is not long enough to show the entire unique identifier. To change this field size, hit ‘L’:

```
11:12:36am up 31 days 22:50, 228 worlds; CPU load average: 0.30, 0.30, 0.32
Change the name field size:  36
DEVICE              PATH/WORLD/PARTITION DQLEN WQLEN ACTV QUED %USD  LOAD   CMDS/s  READS/s WRITES/s MBREAD/s MBWRTN/s
mpx.vmhba3:C0:T            -                1     -    0    0    0  0.00     0.00     0.00     0.00     0.00     0.00
naa.5000c5000bf            -               64     -    0    0    0  0.00     0.00     0.00     0.00     0.00     0.00
naa.5000c5000bf            -               64     -    0    0    0  0.00     0.00     0.00     0.00     0.00     0.00
naa.60060480000            -               32     -    0    0    0  0.00     0.00     0.00     0.00     0.00     0.00
naa.60060480000            -               32     -    1    0    3  0.03     0.00     0.00     0.00     0.00     0.00
{NFS}Burlington            -                -     -    0    -    -     -     0.00     0.00     0.00     0.00     0.00
{NFS}vmlibrary             -                -     -    0    -    -     -     0.00     0.00     0.00     0.00     0.00
```

ESXTop

- Expanding this field now shows the entire unique ID for the LUN and NFS datastores:

```
11:14:55am up 31 days 22:52, 227 worlds; CPU load average: 0.30, 0.30, 0.33

DEVICE                                PATH/WORLD/PARTITION DQLEN WQLEN ACTV QUED %USD  LOAD   CMDS/s  READS/s WRITE
mpx.vmhba3:C0:T0:L0                                  -        1      -    0    0    0  0.00     0.95     0.00     0
naa.5000c5000bf5d68b                                 -       64      -    0    0    0  0.00     0.00     0.00     0
naa.5000c5000bf614af                                 -       64      -    0    0    0  0.00     2.29     0.00     2
naa.600604800001903009195330303031 32                -       32      -    0    0    0  0.00     0.00     0.00     0
naa.600604800001903009195330303031 34                -       32      -    0    0    0  0.00    10.11     4.58     5
{NFS}Burlington ISO Repository                       -        -      -    0    -    -     -     0.00     0.00     0
{NFS}vmlibrary                                       -        -      -    0    -    -     -     0.00     0.00     0
```

- Just as before, we should display only the fields we need to be concerned with by hitting ‘f’:

```
* A:  DEVICE = Device Name
* B:  ID = Path/World/Partition Id
  C:  NUM = Num of Objects
  D:  SHARES = Shares
  E:  BLKSZ = Block Size (bytes)
  F:  QSTATS = Queue Stats
  G:  IOSTATS = I/O Stats
  H:  RESVSTATS = Reserve Stats
* I:  LATSTATS/cmd = Overall Latency Stats (ms)
  J:  LATSTATS/rd = Read Latency Stats (ms)
  K:  LATSTATS/wr = Write Latency Stats (ms)
  L:  ERRSTATS/s = Error Stats
  M:  PAESTATS/s = PAE Stats
  N:  SPLTSTATS/s = SPLIT Stats
  O:  VAAISTATS= VAAI Stats
  P:  VAAILATSTATS/cmd = VAAI Latency Stats (ms)
```

ESXTop

- This view now shows us similar output to the ‘d’ view except we see the statistics for a LUN (all paths).

```
11:19:51am up 31 days 22:57, 228 worlds; CPU load average: 0.41, 0.36, 0.32

DEVICE                                    PATH/WORLD/PARTITION DAVG/cmd KAVG/cmd GAVG/cmd QAVG/cmd
mpx.vmhba3:C0:T0:L0                                  -             1.53     0.03     1.56     0.01
naa.5000c5000bf5d68b                                 -             0.00     0.00     0.00     0.00
naa.5000c5000bf614af                                 -            16.37     0.02    16.39     0.01
naa.600604800001903009195330303031 32                -             0.00     0.00     0.00     0.00
naa.600604800001903009195330303031 34                -             0.31     0.03     0.35     0.01
{NFS}Burlington ISO Repository                       -                -        -     0.00        -
{NFS}vmlibrary                                       -                -        -     0.00        -
```

*Note: The reason there are no statistics for the NFS mounts is due to the fact that these are ISO and template repositories that are not in use currently by this host.

Capturing ESXTop Data

- **If it is required, esxtop data can be collected (referred to as performance snapshot) for later review, either by the user or VMware Global Support Service.**
- **Since this operation is command line driven, it can also be scheduled through 'crond', if performance snapshots are required for a specific period of time (ie: Performance issues seen everyday at 3am)**
- **To capture esxtop data, there are two options:**
  - Use 'vm-support -S' to capture data to playback with "esxtop -R"
  - Use esxtop in batch mode: 'esxtop -b -d <time> -n <iterations>'
- **For instructions on how to schedule with crond, see KB 1033346:**

http://kb.vmware.com/kb/1033346

Capturing ESXTop Data

- Here is an example for collecting esxtop statistics at 2 second intervals, over a 20 second time period:
- # vm-support –S –i 2 –d 20

```
VMware ESX Support Script 1.33


Taking performance snapshots.  This will take about 20 seconds.

Starting detailed scheduler stats.

Starting vscsiStats.
Snapping 0: 20  seconds left.
Done.  1 snapshots created.

Stopping detailed scheduler stats

Stopping vscsiStats.
Done with performance snapshots.
Preparing files: /
Waiting up to 300 seconds for background commands to complete:

Waiting for background commands: -
Creating tar archive ...


File: '/root/esx-2010-12-02--11.40.11238.tgz'
Please attach this file when submitting an incident report.
To file a support incident, go to http://www.vmware.com/support/sr/sr_login.jsp


To see the files collected, run: tar -tzf '/root/esx-2010-12-02--11.40.11238.tgz'
```

ESXTop Replay Mode

- To replay that capture data, extract the compressed file, change to the ‘snaphots’ directory, run the uncompress script, then run ‘esxtop’ in replay mode ‘-R’:

```
[root@bs-tse-i142 ~]# tar -zxf esx-2010-12-02--12.31.23720.tgz
[root@bs-tse-i142 ~]# cd vm-support-bs-tse-i142-2010-12-02--12.31.23720/snapshots/
[root@bs-tse-i142 snapshots]# ./untar.sh
[root@bs-tse-i142 snapshots]# cd ..
[root@bs-tse-i142 vm-support-bs-tse-i142-2010-12-02--12.31.23720]# esxtop -R .
```

```
PCPU USED(%):  92  32  36  38  36  36  36  36 AVG:  43
PCPU UTIL(%):  92  33  38  39  37  38  37  37 AVG:  44
CCPU(%):   6 us,  84 sy,   7 id,   3 wa ;       cs/sec:   2139

    ID    GID NAME             NWLD   %USED    %RUN    %SYS   %WAIT    %RDY   %IDLE  %OVRLP   %CSTP  %MLMTD  %SWPW
     1      1 idle                8  455.01  800.00    0.00    0.00  800.00    0.00    5.65    0.00    0.00    0.0
    11     11 console             1   89.12   89.36    0.18    9.24    1.65    9.16    0.35    0.00    0.00    0.0
   285    285 darmstrong-esx3     6   63.70   64.03    0.03  574.84    1.70  147.96    0.47    0.00    0.00    0.0
   188    188 jayers-server20     9   49.67   49.04    0.01  892.09   18.50  572.09    1.05    0.00    0.00    0.0
   287    287 darmstrong-esx5     5   36.50   36.68    0.02  495.11    2.04  175.72    0.35    0.00    0.00    0.0
   281    281 jchudak-win2k8r     5   25.98   25.84    0.05  500.00    2.47  183.84    0.42    0.00    0.00    0.0
   216    216 jdias-ESX4.0GA-     5   17.60   17.50    0.03  500.00    1.76  194.27    0.23    0.00    0.00    0.0
   276    276 baileym-esxi4       5   14.53   14.49    0.06  500.00    1.45  197.69    0.25    0.00    0.00    0.0
   242    242 XPPro11gR1 (2)      7    9.73    9.69    0.01  700.00    1.09  416.44    0.12    0.01    0.00    0.0
    92     92 CentOs 5.2-32       5    8.09    8.02    0.02  500.00    1.21   96.60    0.15    0.00    0.00    0.0
```

Capturing ESXTop Data

- Running ESXTop in batch mode will generate a .csv (comma separated value) file that will contain statistical counters for every default field. To only capture required fields, you can use a configuration file for esxtop. We will cover this more on the next slide.
- Here is an example for collecting esxtop statistics, in batch mode: 2 second intervals, with 10 iterations:
- # esxtop -b -d 20 -n 100 > hostname_esxtop.csv

```
# esxtop -b -d 2 -n 10 > hostname_esxtop.csv
#
```

Creating ESXTop Configuration File

- **To create a custom configuration file for esxtop to use, follow these steps:**
  - Run ‘esxtop’ without any flags
  - Select which fields you want to display/hide
  - Save the configuration by hitting “W” and then entering a filename
- **Once the configuration file has been save, it can be loaded with the command line parameters:**
- **# esxtop -b -d 2 -n 10 -c esxtoptestrc > hostname_latencyonly.csv**

```
# esxtop -b -d 2 -n 10 -c esxtoptestrc > hostname_latencyonly.csv
#
```

DAVG/KAVG/GAVG/QAVG

- **What is the DAVG?**
  - The DAVG field is the Device Average, which is the amount of time it takes for a SCSI command to leave the HBA, hit the array, and return completed. This is the most effective method of determining if the performance issue being experienced is at the physical layer (switch/array).
- **What are acceptable DAVG values?**
  - Optimal, sustained DAVG latency values would be between **0-5ms** for 4/8GB FC and 10GB iSCSI/NFS. Seeing **5-10ms** latency for 1/2GB FC or 1GB iSCSI/NFS is also acceptable.
  - Seeing latency values of **10-20ms** could start to show some minor performance issues inside the VM. while values of **20-50ms** would show noticeable to significant performance issues. Latency values of **50ms or higher** would make the VMs almost unusable.
- **How should I proceed when my DAVG is sustained at a high level?**
  - Engage the storage team immediately. They may already be aware of the latency due to increased load on the storage array (LUN replication, backups)

DAVG/KAVG/GAVG/QAVG

- **What is the KAVG?**
  - The KAVG field is the Kernel Average, which is the amount of time a SCSI command spends in the vmkernel.
- **What are acceptable KAVG values?**
  - The KAVG should always be less than 1ms.
- **How should I proceed when my KAVG is sustained at a high level?**
  - While rare to see the KAVG value high at all, it has been observed when the queue has been throttled back on the ESX host due to a TASK_SET_FULL condition on the array (KB 1008113), or the queue depth has been reduced on the HBAs due to configuration restrictions (KB 1006001). It has also been observed temporarily on path failover. This is due to commands queuing during the path activation process.

DAVG/KAVG/GAVG/QAVG

- **What is the GAVG?**
  - The GAVG field is the Device Average + Kernel Average + Queue Average, which is the latency perceived by the Guest OS or VM. The GAVG is the only way to measure NFS latency (ESX 4.1)
- **What are acceptable GAVG values?**
  - As the KAVG and QAVG should always be less than 1ms, acceptable GAVG latency values would be the same as acceptable DAVG values.
- **How should I proceed when my GAVG is sustained at a high level?**
  - The same steps should be followed for high DAVG values.

DAVG/KAVG/GAVG/QAVG

- **What is the QAVG?**
  - The QAVG field is the Queue Average, which is the amount of time the SCSI command spends in the HBA driver.
- **What are acceptable QAVG values?**
  - The QAVG should always be less than 1ms.
- **How should I proceed when my QAVG is sustained at a high level?**
  - Just like the KAVG, the QAVG should not have a high latency value, unless there is a legitimate reason for commands to remained queued on the host side.

Queue Depth and ESXTop

- Many people believe that increasing the queue depth will solve a performance issue or improve performance overall, however this can have the opposite effect (KB 1006001 & 1008113).
- The queue depth should only be increased if the array vendor recommends to do so or the queue depth is getting exhausted on the host side.
- As seen in a previous slide, we can observe the queue depth usage and activity:

```
 2:41:43pm up 32 days  2:19, 239 worlds; CPU load average: 0.32, 0.34, 0.34
```

DEVICE	DQLEN	WQLEN	ACTV	QUED	%USD	LOAD	CMDS/s	READS/s	WRITES/s	MBREAD/s	MBWRTN/s
mpx.vmhba3:C0:T0:L0	1	-	0	0	0	0.00	0.57	0.00	0.00	0.00	0.00
naa.5000c5000bf5d68b	64	-	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00
naa.5000c5000bf614af	32	-	32	34	100	2.06	4450.61	4400.44	50.16	73.33	0.80
naa.600604800001903009195330303031 32	32	-	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00
naa.600604800001903009195330303031 34	32	-	32	1	100	1.03	858.50	380.90	476.07	23.67	47.67
{NFS}Burlington ISO Repository	-	-	0	-	-	-	0.00	0.00	0.00	0.00	0.00
{NFS}vmlibrary	-	-	0	-	-	-	0.00	0.00	0.00	0.00	0.00

Agenda

- vCenter Performance Charts, ESXTop & ESXPlot
- **SCSI Reservations**
- Multipathing Considerations

SCSI Reservations

What are SCSI Reservations?

- SCSI Reservations refer to the act of an initiator to using SCSI-2 commands 0x16 (RESERVE) and 0x17 (RELEASE) to lock a LUN for a specific operation.
- While a reservation is placed on a LUN, if another initiator attempts to perform any command on the reserved LUN other than an INQUIRY, REQUEST SENSE, or a PREVENT/ALLOW MEDIUM REMOVAL command the command shall be rejected with RESERVATION CONFLICT device status.
- A reservation may only be released by the initiator that placed the reservation. A release command sent by another initiator will be ignored.
- Any initiator can clear the reservation by issuing a “BUS DEVICE RESET” command or a hard “RESET”.

SCSI Reservations

- Since VMFS-3 is basically a clustered file system, allowing simultaneous access from multiple ESX servers, we need a way of preserving the integrity of the file system when more than one host is updating it.
- VMFS-3 implements a locking protocol to prevent VMware cluster-aware applications from powering on (or otherwise sharing) the contents of a given virtual disk on more than one host at any given time.
- Part of this locking protocol is based on the notion of on-disk locks that protect the metadata on the volume.
- An ESX/ESXi host interested in locked access to metadata must atomically check the owner and lock-state fields in the lock, and if the lock is free, acquire it by writing out its own ID and lock state into the owner and lock-state fields.

SCSI Reservations

- The lock manager currently uses SCSI reservations to check the lock information, establish itself as the owner by writing to the relevant lock fields on disk if the lock is free, and then releases the SCSI reservation.
- ESX uses SCSI-2 non-persistent reservations which implies that only a single host can reserve the LUN in question at any one time. A reboot of that host will clear the reservation on the LUN (non-persistent).
- ESX 4.0 introduces limited support for SCSI-3 persistent reservations to work with Microsoft Windows 2008 Clustering, however the LSISAS controller must be used for the VMs. Only Windows 2k8 is supported with this controller.
- VMFS lock operations are still implemented using SCSI-2 non-persistent reservations, even in ESX 4.x.

A typical SCSI Reservation Conflict message

```
0:12:44:32.598 cpu4:1046)SCSI: vm 1046: 5509: Sync CR at 64
0:12:44:33.520 cpu4:1046)SCSI: vm 1046: 5509: Sync CR at 48
0:12:44:34.512 cpu4:1046)SCSI: vm 1046: 5509: Sync CR at 32
0:12:44:35.482 cpu4:1046)SCSI: vm 1046: 5509: Sync CR at 16
0:12:44:36.490 cpu2:1046)SCSI: vm 1046: 5509: Sync CR at 0
0:12:44:36.490 cpu2:1046)WARNING: SCSI: 5519: Failing I/O due
 to too many reservation conflicts
0:12:44:36.490 cpu2:1046)WARNING: SCSI: 5615: status SCSI
 reservation conflict, rstatus 0xc0de01 for vmhba2:0:3.
 residual R 919, CR 0, ER 3
0:12:44:36.490 cpu2:1046)FSS: 343: Failed with status 0xbad0022
 for f530 28 2 45378334 5f8825b2 17007a7d 1d624ca4 4 1 0 0 0 0
 0
```

"CR" stands for conflict retry. In this case an I/O is being retried due to reservation conflicts. The number at the end of the log statement is the number of retries left. In this case, all retries were exhausted so the I/O failed. In ESX 4.x, this value counts down from 992, and in ESX 4.1 U1, the Sync CR messages are suppressed altogether.

What causes a SCSI Reservation?

- Administrative operations, such as creating or deleting a virtual disk, extending a VMFS volume, or creating or deleting snapshots, result in metadata updates to the file system using locks, and thus result in SCSI reservations.
- Reservations are also generated when you expand a virtual disk, when a snapshot for a virtual machine disk increases in size, or when a thin provisioned virtual disk grows.
- VMotion also use SCSI Reservations, with a reservation placed first by the source ESX, which is subsequently released, and then the destination ESX places a SCSI Reservation on the LUN.
- Note that SCSI reservations are used to acquire or release a lock on a file but is not required when updating host heartbeats to maintain a lock on a file.

For more information, see KB 1005009: http://kb.vmware.com/kb/1005009

What Should I do when I have SCSI Reservation Conflicts?

- **Sometimes SCSI Reservation Conflicts are temporary and my subside. If the SCSI Reservation Conflicts do not subside, issue a LUN "RESET" to the LUN to see if this resolves them. This can be done from an ESX host by issuing the following command:**

```
# vmkfstools -L lunreset /vmfs/devices/disks/naa.xxxxxxxxxxxxxxx
```

***NOTE: To be safe, issue the LUN reset command at least twice.**

- **If the command was successful, you should see messages in the vmkernel similar to the following:**

```
cpu1:1057)<6>scsi(2:0:3:73): DEVICE RESET SUCCEEDED.
```

- **If a LUN reset resolves the issue then this points to a lost reservation as being the root cause of the problem. A lost reservation means that an initiator placed the reservation however it was not able to send the release (HBA goes into fatal error state) or the release was dropped at some level yet a SUCCESS was returned to the initiator (blade environment).**

What Should I do when I have SCSI Reservation Conflicts?

- **If the LUN reset does not resolve the SCSI reservation conflict, the following scenarios may apply:**
  - Array is overloaded:
    - Write pending at 100%
    - Read/write cache exhausted
    - Cache constantly destaging down to disk
    - Array controller/port queue is full
    - LUN Replication/Backups too demanding
  - Host mode setting incorrect for initiator record on the array
  - Array software/firmware bug (LUN flag setting, Storage pool leak, etc)**
  - Hardware Management Agents (HP Insight Manager) are locking the LUN
  - The LUN reporting the SCSI reservation is presented to a non-ESX host
  - The LUN reporting the SCSI reservation is an RDM LUN used by a VM

***Remember: A SCSI reservation conflict is a *symptom*, not the problem.**

Best Practices to avoid SCSI Reservation Conflicts

- Keep HBA, Blade Switch (if applicable), Fabric Switch, and Array firmware up to date. Firmware updates contain fixes!
- Determine if any operations are already happening on the LUN you wish to start another operation that may cause a SCSI Reservation.
- Spread out VCB Backups, VMotion operations, Template deployments, etc.
- Choose one ESX Server as your deployment server to avoid conflicts with multiple ESX servers trying to deploy templates.
- Limit access Administrative operations in vCenter so that you control who can enact an operation that could lead to potential reservations issues (snapshots).

Best Practices to avoid SCSI Reservation Conflicts

- Avoid boot storms by scheduling VM reboots so that there is only one reboot per LUN. In the case of VMware View desktops, choose to leave these desktops always powered on.
- Use care when scheduling backups, antivirus, or LUN replication.
- Ensure that the storage array will not be overloaded with operations that could impact hosts connected.
- Verify that there is no other SCSI Reservations operation happening.
- Use the ATS (Atomic Test and Set) feature of VAAI, if available. More on this feature on the next slide.

Goodbye SCSI Reservations, Hello ATS!

- In vSphere 4.1 the following three offload primitives are supported – Hardware Assisted Locking, Full Copy, and Block Zero. These primitives are also known as Atomic Test and Set, Clone Blocks, and Write Same respectively. We will talk about Clone Blocks and Write Same in later slides.
- Atomic Test and Set (ATS) primitive atomically modifies a sector on disk without the use of SCSI reservations and needing to lock out other hosts from concurrent LUN access. This primitive also reduces the number of commands required to successfully acquire an on-disk lock.
- Upon receiving an ATS command, the array should atomically check if the contents of the disk block at the specified logical block number are the same as initiator-provided existing value, and if yes, replace it with initiator-provided new value.

Goodbye SCSI Reservations, Hello ATS!

- **Unlike SCSI reservations, ATS commands from other hosts should not be rejected with a device status of RESERVATION CONFLICT in the case of in-flight ATS commands from other hosts. Instead, these commands are queued as needed and processed in the same order.**
- **The ATS primitive can significantly improve I/O throughput from guest OS applications to a VMFS volume in the presence of metadata operations, and the number of concurrent cluster-aware VM operations that can be supported in the presence of these I/O intensive enterprise workloads.**
- **ATS will be used for the following operations:**
  - Every lock operation executed by the VMFS-3 lock manager (metadata)
  - Power state operations (power on/off/checkpoint/resume, snapshot and consolidate) of VMs
  - Cluster-wide operations like Storage vMotion, vMotion, DRS

Requirements for VAAI

- **The following are required for VAAI:**
  - ESX/ESXi 4.1
  - Array firmware that supports VAAI primitives (contact array vendor for supportability information)
- **VAAI uses the following SCSI commands:**
  - ATS uses SCSI command 0x89 (COMPARE AND WRITE)
  - Clone Blocks/Full Copy uses SCSI command 0x83 (EXTENDED COPY)
  - Zero Blocks/Write Same uses SCSI command 0x93 (WRITE SAME)
- **These VAAI operations are controlled by the following advanced settings:**
  - /VMFS3/HardwareAcceleratedLocking
  - /DataMover/HardwareAcceleratedMove
  - /DataMover/HardwareAcceleratedInit

Known Issues for VAAI

- **VAAI is enabled by default, and the primitives will issue the specific SCSI commands to determine whether the LUN supports the commands or not. If the commands are not successful, the LUN will be marked as non-VAAI capable.**
- **The reason VAAI is enabled by default is due to NDU (non-disruptive upgrades) or initiator setting changes that could make the LUNs support VAAI.**
- **This default behavior has caused issues for some arrays or fabric configurations that do not support the SCSI commands VAAI uses. Examples of this are the following:**
  - Outdated/unsupported array firmware does not handle SCSI commands correctly and returns unexpected responses
  - Cisco SANTAP controllers do not support the commands and crash

Known Issues for VAAI

- **To check VAAI status, run the command:**

```
# esxcfg-scsidevs -l | egrep "Display Name:|VAAI Status:“
```

- **In the event that your array or fabric configuration does not support VAAI commands and is behaving oddly, disable the VAAI functions on the ESX hosts until a firmware update is available from the vendor.**
- **To disable the VAAI primitives, execute the following commands on each ESX host:**

```
# esxcfg-advcfg -s 0 /DataMover/HardwareAcceleratedMove
# esxcfg-advcfg -s 0 /DataMover/HardwareAcceleratedInit
# esxcfg-advcfg -s 0 /VMFS3/HardwareAcceleratedLocking
```

Agenda

- vCenter Performance Charts, ESXTop & ESXPlot
- SCSI Reservations
- **Multipathing Considerations**

Multipathing Considerations

MRU (Most Recently Used) Path Selection Policy

- The MRU Path Selection Policy was designed for use with Active/Passive or Passive Not Ready (PNR) arrays. These arrays have a sense of LUN ownership, which means a LUN can only be accessed via one controller, not both. As such, Active/Passive arrays have far more failover conditions.
- The MRU policy selects the first working path discovered at system boot time. This would be the first HBA and first path discovered. This design has an inherit flaw as it means that unless failover has occurred, all I/O to the LUNs will be going through the first HBA.
- If this path becomes unavailable, the ESX host switches to an alternative path and continues to use the new path while it is available. It does this because forcing the path back to the other array controller would cause the LUN to change ownership on the array.

Fixed Path Selection Policy

- Originally, the Fixed Path Selection Policy was primarily used by Active/Active arrays however this policy is now used by most ALUA compatible arrays as well as iSCSI arrays with a virtual port.
- The Fixed Path Policy uses the concept of a preferred path, which means that you can select which path you want to use for a LUN and that path will always be used unless a failover scenario occurs. Once the path is restored, however, the storage stack will automatically switch back to the preferred path.
- Fixed was used in the ESX 2.x – 3.x days to load balance I/O from an ESX perspective.
- It is acceptable to use the MRU policy for Active/Active arrays since both controllers can serve up a LUN, however it is NOT acceptable or supported to use Fixed policy on an Active/Passive or Passive Not Ready (PNR) array as this will cause path thrashing to occur.

RoundRobin Path Selection Policy

- **The RoundRobin Path Selection Policy is the only in-box policy that load balances I/O (Fixed doesn't count).**
- **This policy will only use "Active", "Optimized" paths by default as valid paths to switch to. No "Standby" paths will be used as this would cause a trespassing effect/LUN ownership change on Active/Passive or Passive Not Ready (PNR) arrays to occur. This is not a problem for Active/Active arrays since they have no "Standby" paths.**
- **The load balancing mechanism will switch paths by one of two methods:**
  - IOPS threshold (commands sent)
  - throughput threshold (bytes transferred).

RoundRobin Path Selection Policy

- When setting the RoundRobin policy, the default value for IOPS will be 1000 commands and the default throughput threshold will be 10485760 bytes.
- VMware does not recommend settings for either the IOPS or the bytes value with Roundrobin. For information on the correct settings for RoundRobin use with a particular storage array, the storage array vendor MUST be contacted.
- The array vendor will know what settings are optimal for use but they will also know what settings NOT to use.
- Some vendors recommend an ‘iops’ value of ‘1’ to be used however this same setting on another array can have detrimental effects resulting in poor performance or even a possible production outage.
- The RoundRobin policy cannot be used on MSCS Quorum RDM LUNs.

Vendor Recommend Round Robin Settings

- **EMC DMX:**
  http://www.emc.com/collateral/hardware/white-papers/h6531-using-vmware-vsphere-with-emc-symmetrix-wp.pdf
- **EMC Celerra:**
  http://www.emc.com/collateral/software/technical-documentation/h5536-vmware-esx-srvr-using-emc-celerra-stor-sys-wp.pdf
- **NetApp:**
  http://kb.vmware.com/kb/1010713
  http://media.netapp.com/documents/tr-3749.pdf
- **HP EVA:**
  http://h20195.www2.hp.com/v2/GetPDF.aspx/4AA1-2185ENW.pdf

Setting Round Robin Settings From CLI

- **Setting an entire SATP (Storage Array Type Plugin) to Round Robin:**
```
# esxcli nmp satp setdefaultpsp -s VMW_SATP_LSI -p
VMW_PSP_RR
```
- **Setting the Round Robin PSP setting for a single LUN:**
```
# esxcli nmp device setpolicy -d
naa.600174d0010000000010f003048318438 --psp VMW_PSP_RR
```
- **Command line option to set iops value:**
```
# esxcli nmp roundrobin setconfig -d
naa.600174d0010000000010f003048318438 --iops 10 --type
iops
```
- **Command line option to set bytes value:**
```
# esxcli nmp roundrobin setconfig -d
naa.600174d0010000000010f003048318438 --bytes 11 --
type bytes
```

Displaying Round Robin Settings From CLI

- **To displaying Round Robin settings for a LUN:**

```
# esxcli nmp device naa.60a9800050334b356b4a51312f417541

   Device Display Name: NETAPP Fibre Channel Disk
  (naa.60a9800050334b356b4a51312f417541)

   Storage Array Type: VMW_SATP_ALUA

   Storage Array Type Device Config:

{implicit_support=on;explicit_support=off;explicit_allow=on;alua_fol
  lowover=on;{TPG_id=2,TPG_state=AO}{TPG_id=3,TPG_state=ANO}}

   Path Selection Policy: VMW_PSP_RR

   Path Selection Policy Device Config:
  {policy=rr,iops=1000,bytes=10485760,useANO=0;lastPathIndex=3:
  NumIOsPending=0,numBytesPending=0}

   Working Paths: vmhba2:C0:T2:L1, vmhba1:C0:T2:L1
```

Known Issues for Round Robin in ESX 4.x

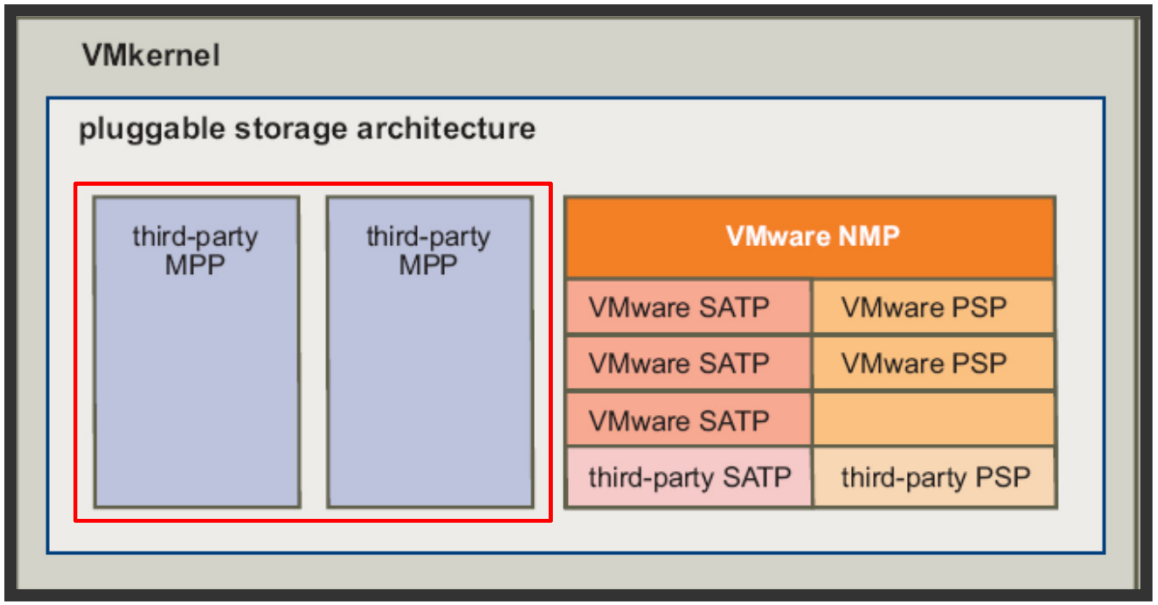
- **There are currently two known issues when using the Round Robin Path Selection Policy:**
  - After setting the IOPS value for roundrobin, the setting is not retained after a reboot and instead shows a much higher value. This is resolved in a patch for ESX 4.0:
    http://kb.vmware.com/kb/1017721
  - The SATP for EMC DMX (VMW_SATP_SYMM) does not distribute I/O evenly with roundrobin after a state change occurs. This state change includes adding a new path (FA) or when losing a paths and recovering. This is also resolved in a patch for ESX 4.0 and ESX 4.1:

    ESX 4.0:
    http://kb.vmware.com/kb/1023759

    ESX 4.1:
    http://kb.vmware.com/kb/1027013

EMC Powerpath for ESX 4.x

- EMC Powerpath was the first third party vendor to make use of vSphere's pluggable storage architecture (PSA).
- The use of Powerpath in vSphere hosts will completely take over the MPP stack. This includes path management, load balancing operations, failover, etc.



EMC Powerpath Features

- **Powerpath/VE 5.1 features include:**
  - Dynamic load balancing (aggregate) and I/O balancing (latency)
  - Auto-Restore of paths
  - Device prioritization
  - Automated performance optimization
  - Dynamic path failover and recovery
  - Monitoring and reporting I/O statistics (all paths, failed commands)
  - Automatic path testing
  - Support for EMC and non-EMC arrays
  - Automatic detection and failover of degraded paths (thresholds met)
  - Additional EMC specific failover codes/conditions (AX4 issue, KB 1029185)

Questions

Lunch

Confidential

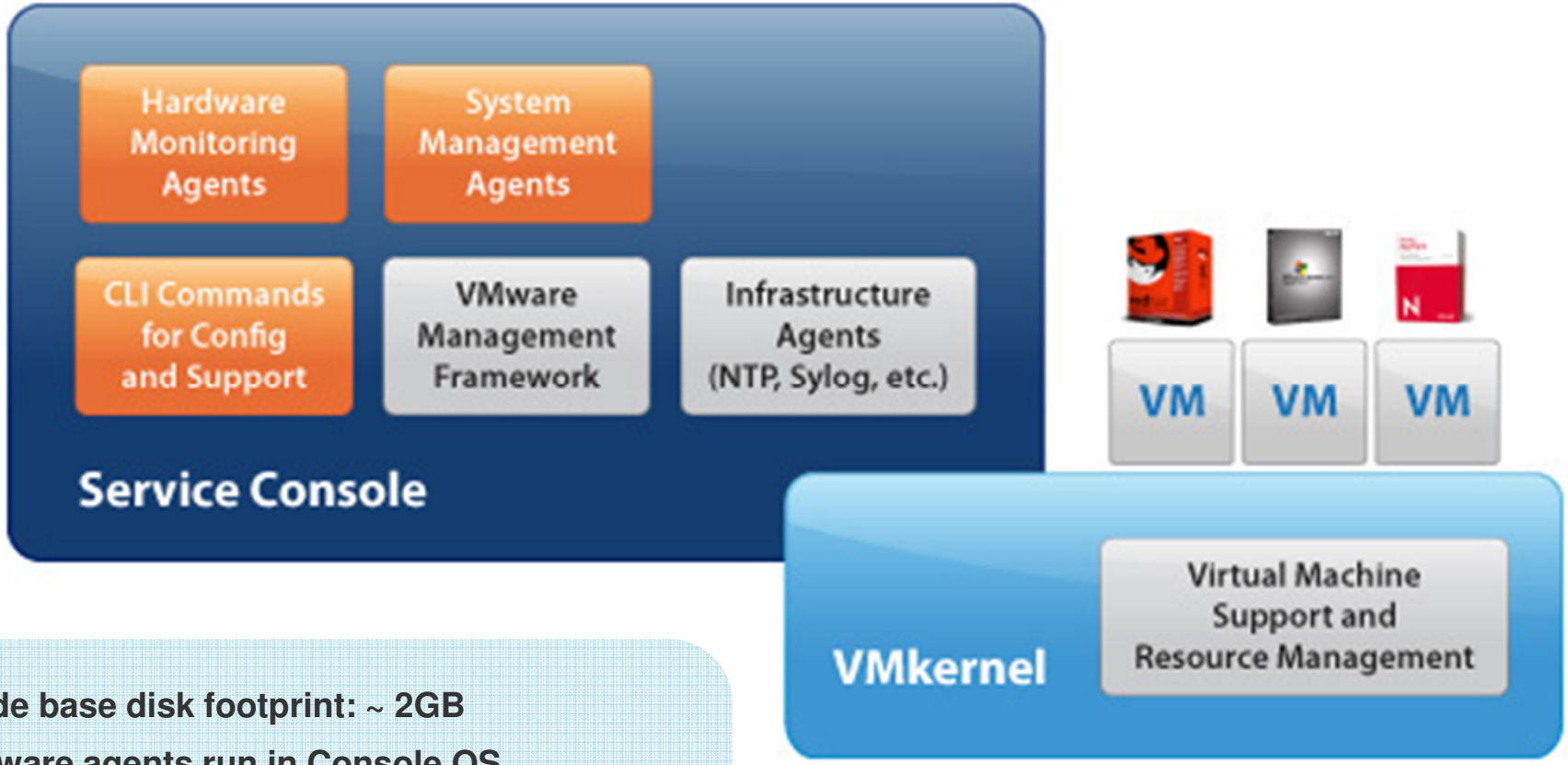
ESXi Readiness

Ben Thomas, Sr. Federal Technical Support Engineer, VMware

What is ESXi?

- Next Generation VMware virtualization platform
- Has been available since version 3.5
- No more bulky service console! (this means fewer patches!)
- Lightweight management interface
- Hypervisor software is the same between ESX and ESXi

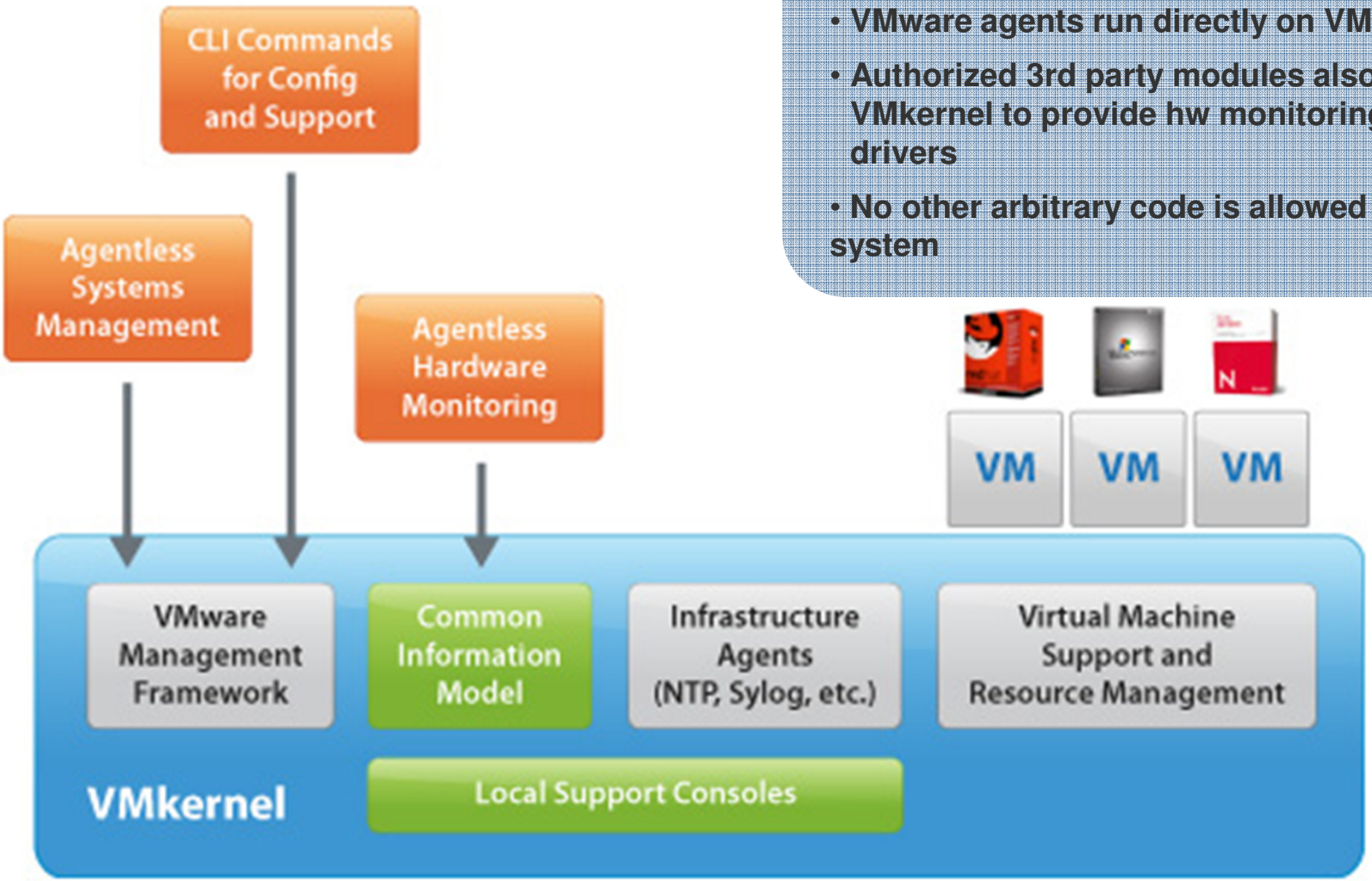
Architecture Comparison - ESX



- Code base disk footprint: ~ 2GB
- VMware agents run in Console OS
- Nearly all other management functionality provided by agents running in the Console OS
- Users must log into Console OS in order to run commands for configuration and diagnostics

Architecture Comparison - ESXi

- Code base disk footprint: <100 MB
- VMware agents run directly on VMkernel
- Authorized 3rd party modules also run in VMkernel to provide hw monitoring and drivers
- No other arbitrary code is allowed on the system



ESXi Myths

- ESXi does not have all of the same features that ESX does
- Different licenses are required for ESXi, existing ESX licenses won't work
- ESXi Software is free software and I do not need a license
- ESXi is harder to manage because it does not have a service console
- ESXi does not support automated (scripted) installs

How do manage it?

- Remotely
  - vCLI/rCLI – Perl scripts
  - PowerCLI – Powershell
  - vMA – vCenter Management Appliance
  - rvc – Ruby vCenter Console (Community)
  - SSH
- Locally
  - DCUI – Menu based management
  - Local Tech Support Mode
- Most ESX related commands that are familiar are available locally and remotely
- Might have to re-tool scripts

Key Points

- ESX and ESXi have virtually the same guest OS compatibility
- ESX and ESXi have virtually the same hardware compatibility
- Able to boot off of certified USB thumb drives
- As of ESXi 4.1 SSH and Local Tech Support Mode are supported
- Much faster to install and to boot

Next Steps

- **Start testing ESXi**
  - ESX and ESXi can comingle, move one host today
- **Ensure 3rd party solutions are ready**
  - Backup and Monitoring software
  - No more agents! (Most agent software cant run on ESXi, this is changing)
- **Become familiar with remote management options**
  - Start migration from any Service Console dependent scripts
  - Become familiar with Powershell and the PowerCLI
- **Plan ESXi migration as a part of next ESX patch cycle**
  - ESXi install time is small and could be fit into a maintenance window

Resources

- ESXi Information Center
- Free “Transition to ESXi” class
- ESXi Migration Guide
- Support Blogs
- VMware Support

Reminder

“VMware would like to remind customers that vSphere 4.1 is the last release to support both the ESX and ESXi hypervisor architecture.

Future major releases will include only the VMware ESXi architecture. For more information visit the ESXi and ESX Info Center.”

Questions

Break

Confidential

Performance Troubleshooting and Best Practices

Ben Thomas, Sr. Federal Technical Support Engineer, VMware

Confidential

Topics

- **Performance Tools**
- **ESXTOP Modes & Common Issues**
  - CPU
  - Memory
  - Network
  - Storage
- **ESXTOP Batch Mode**
- **ESX Plot**

ESX is generic for ESX/ESXi in this presentation

A word about performance…

A performance troubleshooting methodology must provide guidance on how to find the root-cause of the observed performance symptoms, and how to fix the cause once it is found. To do this, it must answer the following questions:

- 1. **How do we know when we are done?**
- 2. Where do we start looking for problems?
- 3. How do we know what to look for to identify a problem?
- 4. How do we find the root-cause of a problem we have identified?
- 5. What do we change to fix the root-cause?
- 6. Where do we look next if no problem is found?

Tools

Measuring Statistics

- ESXTOP / rESXTOP
- ESXTOP Batch Mode
- ESX Plot
- vCenter Performance Graphs

Performance Benchmarking

- cpubusy.vbs
- I/O Meter – Storage
- NetPerf – Network Client/Server testing tool

New Tool – VMware vCenter Operations!

What is ESXTOP?

View real time statistics and health of:

- Hosts
- Virtual Machines
- Memory
- CPU
- DISK
- Network

Similar Tools:

- Linux/Unix Systems
  - top
  - vmstat
  - iostat
- Windows Systems
  - PerfMon

Accessing ESXTOP

2 Ways to access ESXTOP:

1. Directly from the ESX or ESXi Host command line
   - ssh to the host directly or on console

2. rESXTOP (Remote ESXTOP)
   - Available through vMA and RCLI

Anatomy of ESXTOP

CPU View (default)

- Host/VM CPU stats
- CPU Usage Total
- % used per VM / per VCPU
- %RDY (over commitment)

Memory View

- Host/VM memory stats
- Swap stats
- Memory Ballooning stats

Network View

- Host/VM network stats
- Per NIC usage
- Per VM usage
- Per VM and total throughput

Anatomy of ESXTOP

Disk Adapter View

- Per Host Adapter stats
  - Total Commands
  - Latencies
  - Reads/Writes

Disk Device View

- Provides similar statistics as Disk Adapter View but Per LUN/Path (more granular).
- Per LUN Queue Depth usage

VM Disk View

- Total Commands Per VM
- Per VM Reads/Writes
- Per VM latencies

Navigating ESXTOP

Changing Views in ESXTOP:

c → CPU view (default when esxtop starts)

m → Memory view

d → DISK Adapter view

u → Disk Device view

v →VM Disk view

n →Network view

ESXTOP Help

```
root@wdc-tse-i04:~
Secure mode Off

Esxtop: top for ESX

These single-character commands are available:

^L      - redraw screen
space   - update display
h or ?  - help; show this text
q       - quit

Interactive commands are:

fF      Add or remove fields
oO      Change the order of displayed fields
s       Set the delay in seconds between updates
#       Set the number of instances to display
W       Write configuration file ~/.esxtop41rc
k       Kill a world
e       Expand/Rollup Cpu Statistics
V       View only VM instances
L       Change the length of the NAME field
l       Limit display to a single group

Sort by:
        U:%USED         R:%RDY          N:GID
Switch display:
        c:cpu           i:interrupt     m:memory        n:network
        d:disk adapter  u:disk device   v:disk VM       p:power mgmt

Hit any key to continue:
```

Statistic Fields

10.131.1.219 - PuTTY

```
 8:44:35pm up 25 days 22:25, 204 worlds; CPU load average: 0.00, 0.01, 0.01
PCPU USED(%): 0.2 0.1 0.1 0.0 0.1 0.0 0.1 0.0 1.0 0.4 0.8 0.4 6.4 0.2 0.3 0.2 AVG: 0.7
PCPU UTIL(%): 0.4 0.1 0.2 0.1 0.1 0.0 0.1 0.0 1.1 0.5 0.8 0.5 6.2 0.3 0.3 0.2 AVG: 0.7
CORE UTIL(%): 0.5     0.2     0.2     0.2     1.6     1.3     6.4     0.5     AVG: 1.4
```

Fields

ID	GID	NAME	NWLD	%USED	%RUN	%SYS	%WAIT	%RDY	%IDLE	%OVRLP	%CSTP	%MLMTD	%SWPWT
1	1	idle	16	796.65	1600.00	0.00	0.00	1600.00	0.00	0.38	0.00	0.00	0.00
17	17	vmkapimod	6	9.20	8.76	0.00	596.43	0.00	0.00	0.00	0.00	0.00	0.00
1248677	1248677	esxtop.3874982	1	0.47	0.44	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
516742	516742	hostd.1553765	19	0.13	0.04	0.00	1900.00	0.02	0.00	0.00	0.00	0.00	0.00
2	2	system	8	0.12	0.08	0.00	800.00	0.00	0.00	0.00	0.00	0.00	0.00
517095	517095	vpxa.1554157	18	0.11	0.10	0.00	1800.00	0.04	0.00	0.00	0.00	0.00	0.00
517462	517462	sfcb-ProviderMa	4	0.08	0.14	0.00	400.00	0.00	0.00	0.02	0.00	0.00	0.00
7	7	helper	61	0.07	0.01	0.00	6100.00	0.00	0.00	0.00	0.00	0.00	0.00
516786	516786	sensord.1553811	1	0.01	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
517161	517161	vmware-usbarbit	2	0.01	0.02	0.00	200.00	0.00	0.00	0.00	0.00	0.00	0.00
1023202	1023202	dropbearmulti.3	1	0.01	0.01	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
8	8	drivers	10	0.00	0.01	0.00	1000.00	0.00	0.00	0.00	0.00	0.00	0.00
516855	516855	vprobed.1553892	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
517058	517058	openwsmand.9926	3	0.00	0.00	0.00	300.00	0.00	0.00	0.00	0.00	0.00	0.00
517461	517461	sfcb-ProviderMa	6	0.00	0.00	0.00	600.00	0.00	0.00	0.00	0.00	0.00	0.00
516726	516726	ntpd.1553747	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
516809	516809	storageRM.15538	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
258	258	vmklogger.4451	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
644	644	FT	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
9	9	vmotion	4	0.00	0.00	0.00	400.00	0.00	0.00	0.00	0.00	0.00	0.00
516754	516754	sh.1553777	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
1248664	1248664	sh.3874808	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
516764	516764	net-lbt.1553787	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
517151	517151	sh.1554221	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
564262	564262	nssquery.169881	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
516776	516776	sh.1553801	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00

Navigating ESXTOP

Pressing ‘f’ in any View will display available fields.

```
10.131.1.219 - PuTTY

Current Field order: ABcDEFghij

* A:  ID = Id
* B:  GID = Group Id
  C:  LWID = Leader World Id (World Group Id)
* D:  NAME = Name
* E:  NWLD = Num Members
* F:  %STATE TIMES = CPU State Times
  G:  EVENT COUNTS/s = CPU Event Counts
  H:  CPU ALLOC = CPU Allocations
  I:  SUMMARY STATS = CPU Summary Stats
  J:  POWER STATS = CPU Power Stats

Toggle fields with a-j, any other key to return:
```

‘*’ Next to field means that it is enabled and viewable in the main View.

CPU

CPU Related Performance Problems

- **Starved VM** – Too Few vCPUs given to a guest
- **Misconfigured VM** - Improper CPU limit set on guest
- **Bloated VM** - Giving a guest unnecessary CPU resources
- **Misconfigured Host -** Over committing physical CPUs
- %RDY and %CSTP metric

Starved VM

```
root@wdc-tse-i04:~
 8:54:54am up  1:17, 137 worlds; CPU load average: 0.13, 0.04, 0.02
PCPU USED(%): 1.8 0.1  99 0.1 0.3 0.2 0.0 0.1 AVG:  12
PCPU UTIL(%): 1.9 0.1 100 0.1 0.4 0.3 0.1 0.1 AVG:  12
CCPU(%):   0 us,   2 sy,  98 id,   0 wa ;       cs/sec:     57

    ID    GID NAME             NWLD   %USED    %RUN    %SYS   %WAIT    %RDY   %IDLE  %OVRLP   %CST
     1      1 idle                8  697.27  699.78    0.00    0.00  100.24    0.00    0.46    0.0
    57     57 TestVM              4   99.57  100.41    0.01  299.69    0.03    0.00    0.06    0.0
    11     11 console             1    1.74    1.72    0.02   98.25    0.03   98.25    0.00    0.0
     7      7 helper             78    0.03    0.03    0.00 7800.00    0.01    0.00    0.00    0.0
    19     19 vmkapimod          11    0.02    0.02    0.00 1100.00    0.00    0.00    0.00    0.0
     8      8 drivers            10    0.01    0.01    0.00 1000.00    0.00    0.00    0.00    0.0
    55     55 vmkiscsid.4298      2    0.00    0.00    0.00  200.00    0.00    0.00    0.00    0.0
     2      2 system              7    0.00    0.00    0.00  700.00    0.00    0.00    0.00    0.0
    47     47 storageRM.4261      1    0.00    0.00    0.00  100.00    0.00    0.00    0.00    0.0
    49     49 sensord.4264        1    0.00    0.00    0.00  100.00    0.00    0.00    0.00    0.0
     9      9 vmotion             4    0.00    0.00    0.00  400.00    0.00    0.00    0.00    0.0
    46     46 FT                  1    0.00    0.00    0.00  100.00    0.00    0.00    0.00    0.0
    48     48 vobd.4263           6    0.00    0.00    0.00  600.00    0.00    0.00    0.00    0.0
    50     50 net-cdp.4270        1    0.00    0.00    0.00  100.00    0.00    0.00    0.00    0.0
    51     51 net-lbt.4271        1    0.00    0.00    0.00  100.00    0.00    0.00    0.00    0.0
    52     52 vmware-vmkauthd     1    0.00    0.00    0.00  100.00    0.00    0.00    0.00    0.0
```

Low %IDLE

High %USED

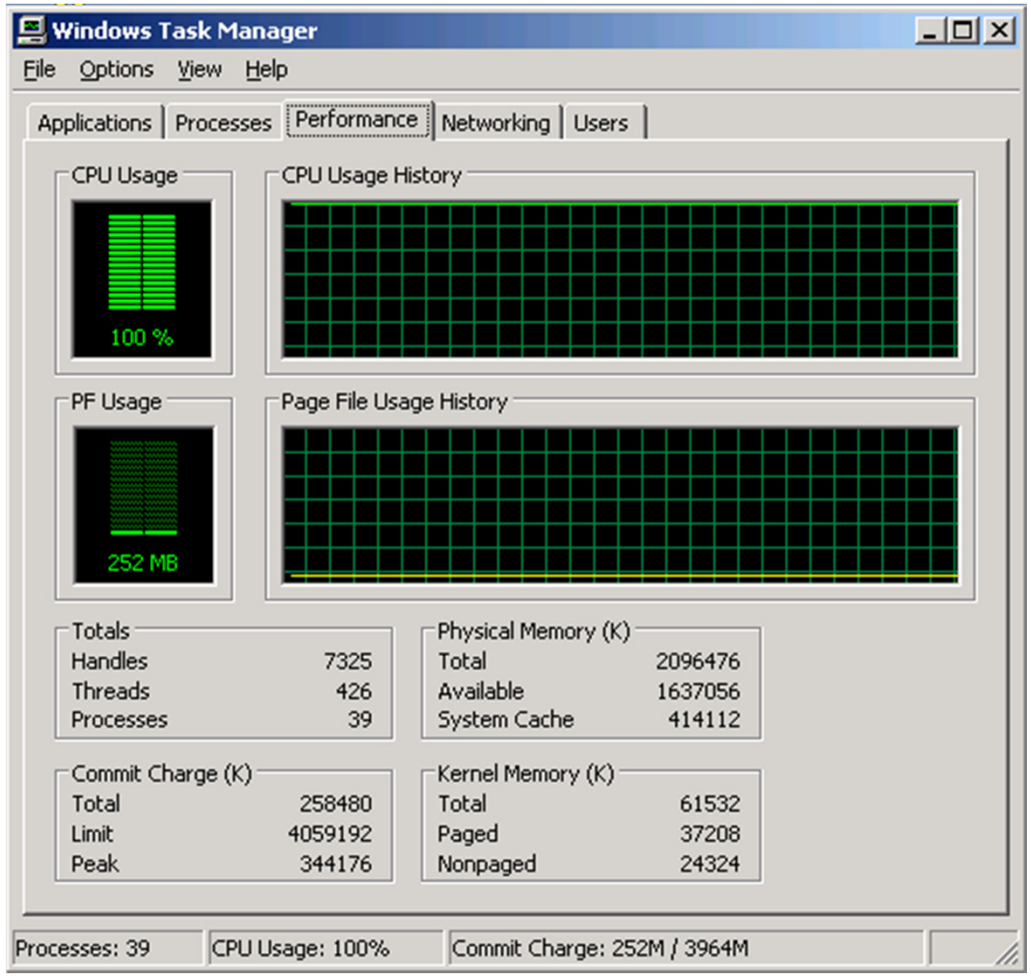
Starved VM

```
root@wdc-tse-i04:~
 8:57:04am up  1:19, 137 worlds; CPU load average: 0.13, 0.10, 0.04
PCPU USED(%): 1.8 0.1 100 0.1 0.1 0.1 0.2 0.2 AVG:  12
PCPU UTIL(%): 1.9 0.1 100 0.2 0.2 0.1 0.4 0.3 AVG:  12
CCPU(%):   0 us,   2 sy,  97 id,   1 wa ;       cs/sec:     88

    ID    GID NAME             NWLD   %USED    %RUN    %SYS   %WAIT    %RDY   %IDLE  %OVRLP   %CST
     1      1 idle                8  697.36  703.31    0.00    0.00   96.68    0.00    0.50    0.0
  4390     57 vmware-vmx          1    0.03    0.04    0.00   99.96    0.00    0.00    0.00    0.0
  4392     57 vmassistant.439     1    0.03    0.04    0.00   99.96    0.00    0.00    0.00    0.0
  4393     57 mks:TestVM          1    0.27    0.30    0.00   99.67    0.02    0.00    0.00    0.0
  4394     57 vcpu-0:TestVM       1   99.96  100.00    0.00    0.00    0.00    0.00    0.05    0.0
    11     11 console             1    1.76    1.74    0.03   98.22    0.03   98.21    0.01    0.0
    19     19 vmkapimod          11    0.05    0.05    0.00 1099.95    0.00    0.00    0.00    0.0
     7      7 helper             78    0.04    0.04    0.00 7800.00    0.01    0.00    0.00    0.0
     8      8 drivers            10    0.01    0.01    0.00 1000.00    0.00    0.00    0.00    0.0
    55     55 vmkiscsid.4298      2    0.00    0.00    0.00  199.99    0.00    0.00    0.00    0.0
     2      2 system              7    0.00    0.00    0.00  700.00    0.00    0.00    0.00    0.0
    49     49 sensord.4264        1    0.00    0.00    0.00  100.00    0.00    0.00    0.00    0.0
    47     47 storageRM.4261      1    0.00    0.00    0.00  100.00    0.00    0.00    0.00    0.0
    50     50 net-cdp.4270        1    0.00    0.00    0.00  100.00    0.00    0.00    0.00    0.0
     9      9 vmotion             4    0.00    0.00    0.00  400.00    0.00    0.00    0.00    0.0
    46     46 FT                  1    0.00    0.00    0.00  100.00    0.00    0.00    0.00    0.0
    48     48 vobd.4263           6    0.00    0.00    0.00  599.99    0.00    0.00    0.00    0.0
    51     51 net-lbt.4271        1    0.00    0.00    0.00  100.00    0.00    0.00    0.00    0.0
    52     52 vmware-vmkauthd     1    0.00    0.00    0.00  100.00    0.00    0.00    0.00    0.0
```

CPU time being used by vCPU Process, may need to add another vCPU.

Misconfigured VM



Misconfigured VM

```
root@wdc-tse-i04:~
 9:02:05am up  1:24, 137 worlds; CPU load average: 0.09, 0.12, 0.07
PCPU USED(%): 2.8 0.1 0.1 0.1 0.1 2.1 0.1 0.2 AVG: 0.7
PCPU UTIL(%): 2.9 0.2 0.2 0.2 0.1 2.1 0.2 0.3 AVG: 0.8
CCPU(%):   0 us,   4 sy,  95 id,   2 wa ;       cs/sec:     97
```

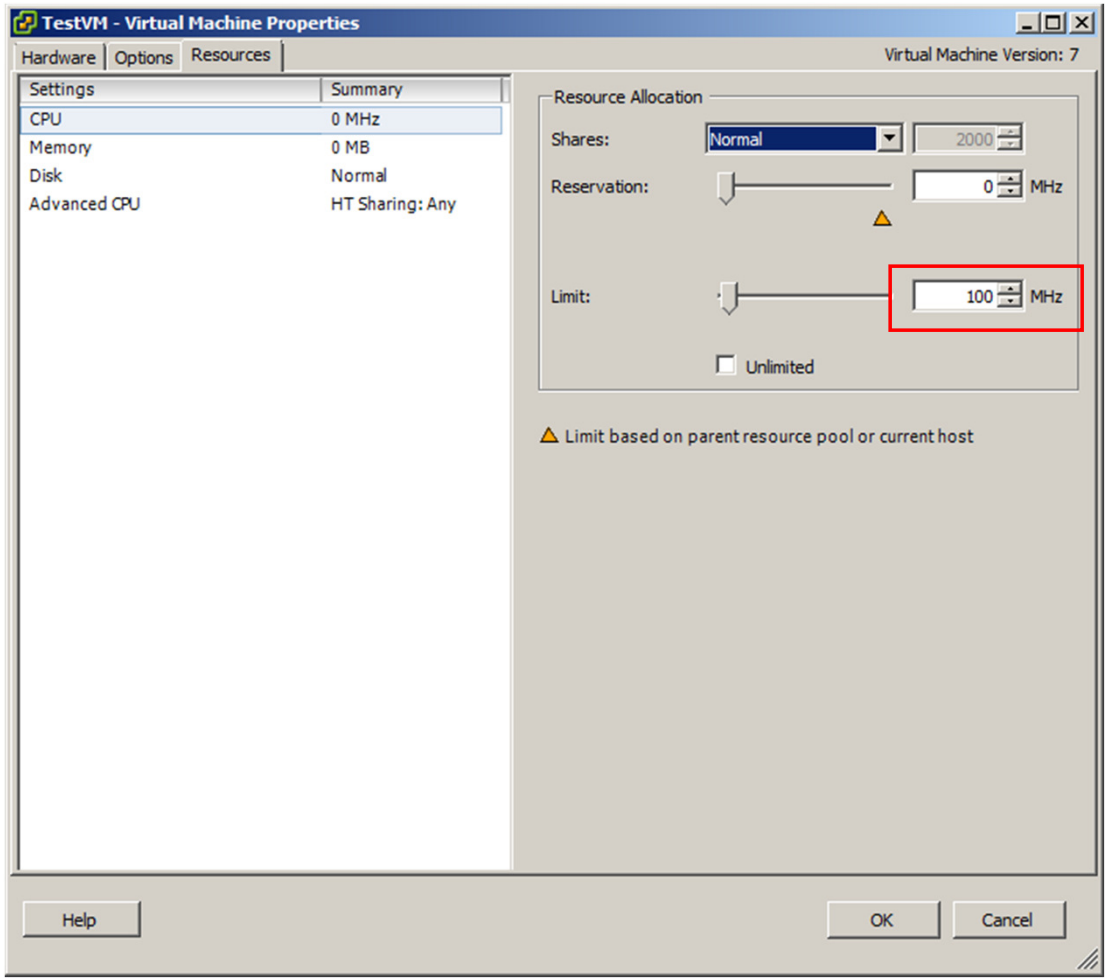
ID	GID	NAME	NWLD	%USED	%RUN	%SYS	%WAIT	%RDY	%IDLE	%OVRLP	%CSTP	%MLMTD	%SWPWT
1	1	idle	8	794.70	800.00	0.00	0.00	800.00	0.00	0.46	0.00	0.00	0.00
11	11	console	1	2.68	2.71	0.03	97.26	0.03	97.25	0.01	0.00	0.00	0.00
57	57	TestVM	4	2.24	2.25	0.01	280.68	117.07	0.00	0.01	0.00	117.05	0.00
7	7	helper	78	0.05	0.05	0.00	7799.89	0.01	0.00	0.00	0.00	0.00	0.00
19	19	vmkapimod	11	0.02	0.02	0.00	1099.97	0.00	0.00	0.00	0.00	0.00	0.00
8	8	drivers	10	0.01	0.01	0.00	999.98	0.00	0.00	0.00	0.00	0.00	0.00
55	55	vmkiscsid.4298	2	0.00	0.01	0.00	199.99	0.00	0.00	0.00	0.00	0.00	0.00
49	49	sensord.4264	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
47	47	storageRM.4261	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
2	2	system	7	0.00	0.00	0.00	699.98	0.00	0.00	0.00	0.00	0.00	0.00
9	9	vmotion	4	0.00	0.00	0.00	400.00	0.00	0.00	0.00	0.00	0.00	0.00
46	46	FT	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
48	48	vobd.4263	6	0.00	0.00	0.00	600.00	0.00	0.00	0.00	0.00	0.00	0.00
50	50	net-cdp.4270	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
51	51	net-lbt.4271	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
52	52	vmware-vmkauthd	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00

Low CPU Usage

High %RDY

High %MLMTD

Misconfigured VM



Bloated VM

```
root@wdc-tse-i04:~
 9:20:51am up  1:43, 140 worlds; CPU load average: 0.13, 0.15, 0.06
PCPU USED(%): 5.7 1.0 1.5 3.2 0.9 0.6  99 0.2 AVG:  14
PCPU UTIL(%): 6.3 1.3 1.7 3.5 1.0 0.7  99 0.4 AVG:  14
CCPU(%):   1 us,   4 sy,  95 id,   0 wa ;       cs/sec:   3192
```

ID	GID	NAME	NWLD	%USED	%RUN	%SYS	%WAIT	%RDY	%IDLE	%OVRLP	%CSTP	%MLMTD	%SWPWT
1	1	idle	8	685.94	800.00	0.00	0.00	800.00	0.00	0.61	0.00	0.00	0.00
58	58	TestVM	7	102.30	103.22	0.03	596.56	0.23	297.03	0.08	0.00	0.00	0.00
11	11	console	1	5.23	5.53	0.02	94.42	0.05	94.27	0.02	0.00	0.00	0.00
19	19	vmkapimod	11	3.31	3.36	0.00	1096.66	0.01	0.00	0.00	0.00	0.00	0.00
7	7	helper	78	0.80	0.88	0.00	7798.91	0.19	0.00	0.01	0.00	0.00	0.00
8	8	drivers	10	0.01	0.01	0.00	1000.00	0.00	0.00	0.00	0.00	0.00	0.00
55	55	vmkiscsid.4298	2	0.01	0.01	0.00	199.99	0.00	0.00	0.00	0.00	0.00	0.00
49	49	sensord.4264	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
47	47	storageRM.4261	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
2	2	system	7	0.00	0.00	0.00	700.00	0.00	0.00	0.00	0.00	0.00	0.00
9	9	vmotion	4	0.00	0.00	0.00	400.00	0.00	0.00	0.00	0.00	0.00	0.00
46	46	FT	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
48	48	vobd.4263	6	0.00	0.00	0.00	600.00	0.00	0.00	0.00	0.00	0.00	0.00
50	50	net-cdp.4270	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
51	51	net-lbt.4271	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
52	52	vmware-vmkauthd	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00

Has CPU Usage

High %IDLE

Bloated VM

```
root@wdc-tse-i04:~
 9:19:36am up  1:42, 140 worlds; CPU load average: 0.15, 0.13, 0.05
PCPU USED(%): 2.7 0.7 0.7 0.5 0.6 0.7  99 0.1 AVG:  13
PCPU UTIL(%): 3.1 0.9 0.9 0.6 0.7 0.8  99 0.1 AVG:  13
CCPU(%):   0 us,   2 sy,  98 id,   0 wa ;       cs/sec:     67
```

ID	GID	NAME	NWLD	%USED	%RUN	%SYS	%WAIT	%RDY	%IDLE	%OVRLP	%CSTP	%MLMTD	%SWPWT
1	1	idle	8	693.95	738.93	0.00	0.00	61.11	0.00	0.53	0.00	0.00	0.00
4411	58	vmware-vmx	1	0.06	0.07	0.00	99.93	0.00	0.00	0.00	0.00	0.00	0.00
4413	58	vmassistant.441	1	0.04	0.05	0.00	99.96	0.00	0.00	0.00	0.00	0.00	0.00
4417	58	mks:TestVM	1	0.31	0.34	0.00	99.63	0.03	0.00	0.01	0.00	0.00	0.00
4418	58	vcpu-0:TestVM	1	1.01	1.11	0.01	98.84	0.05	98.71	0.01	0.00	0.00	0.00
4419	58	vcpu-1:TestVM	1	99.76	99.91	0.00	0.09	0.00	0.00	0.05	0.00	0.00	0.00
4420	58	vcpu-2:TestVM	1	0.97	1.09	0.00	98.85	0.06	98.63	0.01	0.00	0.00	0.00
4421	58	vcpu-3:TestVM	1	0.99	1.02	0.00	98.89	0.09	98.66	0.00	0.00	0.00	0.00
11	11	console	1	1.97	1.99	0.01	97.90	0.11	97.90	0.00	0.00	0.00	0.00
7	7	helper	78	0.03	0.03	0.00	7800.00	0.01	0.00	0.00	0.00	0.00	0.00
19	19	vmkapimod	11	0.02	0.03	0.00	1100.00	0.00	0.00	0.00	0.00	0.00	0.00
8	8	drivers	10	0.01	0.01	0.00	1000.00	0.00	0.00	0.00	0.00	0.00	0.00
55	55	vmkiscsid.4298	2	0.01	0.01	0.00	199.99	0.00	0.00	0.00	0.00	0.00	0.00
49	49	sensord.4264	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
2	2	system	7	0.00	0.00	0.00	700.00	0.00	0.00	0.00	0.00	0.00	0.00
47	47	storageRM.4261	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
9	9	vmotion	4	0.00	0.00	0.00	400.00	0.00	0.00	0.00	0.00	0.00	0.00
46	46	FT	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
48	48	vobd.4263	6	0.00	0.00	0.00	600.00	0.00	0.00	0.00	0.00	0.00	0.00
50	50	net-cdp.4270	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
51	51	net-lbt.4271	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00
52	52	vmware-vmkauthd	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00

MostCPU Usage on 1 vCPU

Most other vCPU are Idle

Bloated VM

Do all your VMs really need to have multiple vCPUs?

- Use a “Least Resources” approach, make VMs prove their needs.

Additional VMkernel memory overhead for each additional vCPUs.

- Don’t forget to include this when sizing hosts.

Guest OS memory overhead for each additional CPU.

- SMP vs Non-SMP Kernel, this is an issue in ALL operating systems.

pCPU != vCPU

%RDY and %CSTP

%RDY – Percent Ready

The percentage of time the world was ready to run but sitting in a run queue waiting for CPU scheduler to let it run on a pCPU

%CSTP – Percent Costop

The percentage of time the CPU scheduler is artificially sleeping a vCPU thread to let fellow threads “catch up”.

This co-deschedule state is only meaningful for SMP VMs. Roughly speaking, ESX CPU scheduler deliberately puts a vCPU in this state, if this vCPU advances much farther than other vCPUs.

Overcommitted Host

root@wdc-tse-i04:~

```
 2:11:52am up 18:34, 244 worlds; CPU load average: 2.83, 2.78, 2.77
PCPU USED(%):  99 100  99  99  99  99  99 100 AVG: 100
PCPU UTIL(%): 100 100 100 100 100 100 100 100 AVG: 100
CCPU(%):   0 us,   3 sy,  96 id,   0 wa ;       cs/sec:     96
```

ID	GID	NAME	NWLD	%USED	%RUN	%SYS	%WAIT	%RDY	%IDLE	%OVRLP	%CSTP	%MLMTD	%SWPW
91	91	OtherVM9	5	36.69	35.98	0.00	387.37	75.72	79.50	0.05	0.87	0.00	0.0
94	94	OtherVM14	5	36.64	36.69	0.00	380.53	82.13	74.58	0.06	0.59	0.00	0.0
100	100	OtherVM19	5	36.56	36.46	0.00	383.55	77.90	76.85	0.05	2.08	0.00	0.0
98	98	OtherVM16	5	36.51	36.58	0.02	386.63	75.75	75.22	0.09	0.99	0.00	0.0
82	82	OtherVM1	5	36.36	36.91	0.01	384.02	73.89	76.97	0.06	5.12	0.00	0.0
86	86	OtherVM6	5	36.34	36.40	0.00	380.54	80.92	66.97	0.06	2.10	0.00	0.0
93	93	OtherVM4	5	36.34	36.16	0.00	384.82	77.18	77.33	0.06	1.79	0.00	0.0
85	85	OtherVM7	5	36.31	36.37	0.00	386.36	73.43	75.99	0.06	3.80	0.00	0.0
89	89	OtherVM10	5	36.23	36.27	0.02	383.70	77.95	71.29	0.06	2.03	0.00	0.0
101	101	TestVM	6	36.21	36.23	0.16	413.63	111.44	0.00	0.19	38.67	0.00	0.0
90	90	OtherVM8	5	36.20	36.20	0.00	386.48	75.18	75.52	0.07	2.08	0.00	0.0
97	97	OtherVM13	5	36.16	36.48	0.00	379.18	83.02	73.67	0.06	1.27	0.00	0.0
95	95	OtherVM18	5	36.10	36.16	0.00	388.83	73.31	78.95	0.06	1.65	0.00	0.0
87	87	OtherVM12	5	36.09	36.13	0.00	389.40	72.91	80.42	0.05	1.51	0.00	0.0
80	80	OtherVM3	5	36.08	35.90	0.01	381.44	81.03	69.09	0.06	1.56	0.00	0.0
99	99	OtherVM20	5	36.08	36.13	0.00	380.46	78.13	69.98	0.05	5.25	0.00	0.0
92	92	OtherVM11	5	36.08	36.14	0.00	388.18	74.34	79.42	0.07	1.28	0.00	0.0
83	83	OtherVM2	5	36.05	36.11	0.00	380.54	80.13	67.52	0.06	3.15	0.00	0.0
88	88	OtherVM15	5	35.99	36.04	0.00	383.65	78.36	76.87	0.05	1.90	0.00	0.0
96	96	OtherVM17	5	35.97	36.10	0.00	382.10	77.44	70.97	0.06	4.31	0.00	0.0
84	84	OtherVM5	5	35.65	35.70	0.00	385.04	78.14	78.41	0.06	1.07	0.00	0.0
81	81	OtherVM0	5	35.44	35.52	0.00	379.21	83.08	65.91	0.05	2.13	0.00	0.0
11	11	console	1	3.44	3.39	0.06	96.50	0.09	96.50	0.01	0.00	0.00	0.0
7	7	helper	78	0.05	0.05	0.00	7798.00	0.14	0.00	0.00	0.00	0.00	0.0
19	19	vmkapimod	11	0.02	0.02	0.00	1099.73	0.00	0.00	0.00	0.00	0.00	0.0
8	8	drivers	10	0.01	0.01	0.00	999.76	0.01	0.00	0.00	0.00	0.00	0.0

Memory

Memory Related Performance Problems

- **ESX Memory Management Stages**
  - Transparent Page Sharing
  - Ballooning
  - Memory Compression (new in ESX 4.1)
  - Host Swapping

- **Important ESX Host Memory Counters**

- **Important VM Memory Counters**

- **Starved VM -** Not enough RAM given to a guest OS
  - Guest OS Swapping

- **Bloated Host -** Over commitment of host memory
  - Host Swapping (BAD)

- **Hardware MMU and Monitor Modes**

ESX Memory Management Stages

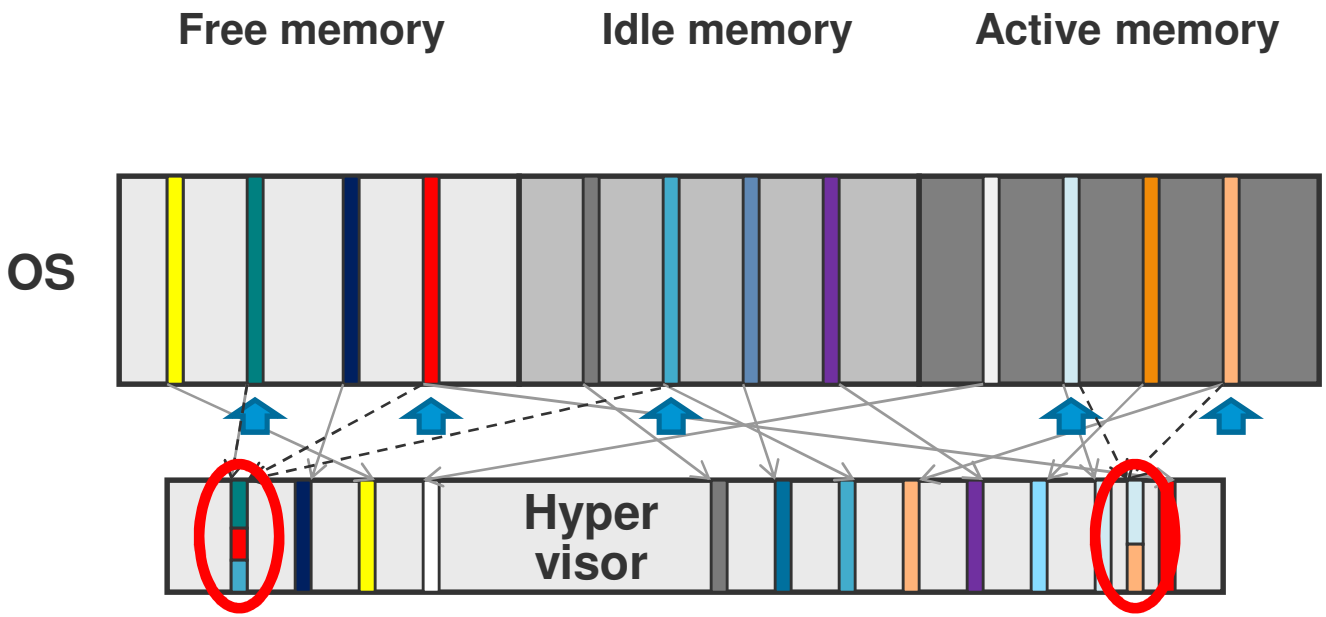
- **Transparent Page Sharing (TPS)**
  - Common Memory used by all VMs on the same host

- **Ballooning**
  - Needs VMware tools
  - Allows ESX to reclaim unused memory pages

- **Memory Compression** (ESX 4.1 only)
  - Think of this as zip for memory

- **Swapping** (ESX Level Swap)
  - Swap memory to .vswp file per VM.

New in vSphere 4.1: Memory Compression



- Solution to disk swap-in problem: try to compress before swapping
  - Decompression is up to 100x faster than swap-in!
- Fall back to swapping if guest memory is uncompressible

Important ESX Host Memory Counters

```
root@wdc-tse-i04:~
 1:23:03pm up  5:45, 138 worlds; MEM overcommit avg: 0.00, 0.07, 0.39
PMEM  /MB: 14334   total:   380     cos,   604 vmk,     744 other,  12605 free
VMKMEM/MB: 13740 managed:   824 minfree,  1779 rsvd,  11961 ursvd,  high state
COSMEM/MB:    53    free:   760  swap_t,   760 swap_f:   0.00 r/s,   0.00 w/s
PSHARE/MB:  1461  shared,    24  common:  1437 saving
SWAP  /MB:     0    curr,     0 rclmtgt:                0.00 r/s,   0.00 w/s
ZIP   /MB:     0  zipped,     0   saved
MEMCTL/MB:     0    curr,     0  target,  1330 max

   GID NAME              MEMSZ    GRANT    SZTGT     TCHD   TCHD_W %ACTV %ACTVS %ACTVF %ACTVN    SWCUR    SWTGT   S
    58 TestVM          2048.00  2048.00   733.80    20.48    20.48     2      0      1      1     0.00     0.00
    55 vmkiscsid.4298    62.81     1.60     1.76     1.60     1.60     0      0      0      0     0.00     0.00
    48 vobd.4263         26.45     4.88     5.36     4.88     4.88     0      0      0      0     0.00     0.00
    47 storageRM.4261    15.36     4.71     5.18     4.71     4.71     0      0      0      0     0.00     0.00
    51 net-lbt.4271      14.46     4.19     4.61     4.19     4.19     0      0      0      0     0.00     0.00
    49 sensord.4264       6.11     1.39     1.53     1.39     1.39     0      0      0      0     0.00     0.00
    52 vmware-vmkauthd    6.10     2.18     2.40     2.18     2.18     0      0      0      0     0.00     0.00
    50 net-cdp.4270       3.54     0.35     0.39     0.35     0.35     0      0      0      0     0.00     0.00
```

Important VM Memory Counters

root@wdc-tse-i04:~

```
 1:23:03pm up  5:45, 138 worlds; MEM overcommit avg: 0.00, 0.07, 0.39
PMEM  /MB: 14334   total:   380     cos,   604 vmk,     744 other,  12605 free
VMKMEM/MB: 13740 managed:   824 minfree,  1779 rsvd,  11961 ursvd,  high state
COSMEM/MB:    53    free:   760  swap_t,   760 swap_f:   0.00 r/s,   0.00 w/s
PSHARE/MB:  1461  shared,    24  common:  1437 saving
SWAP  /MB:     0    curr,     0 rclmtgt:                0.00 r/s,   0.00 w/s
ZIP   /MB:     0  zipped,     0   saved
MEMCTL/MB:     0    curr,     0  target,  1330 max
```

GID	NAME	MEMSZ	GRANT	SZTGT	TCHD	TCHD_W	%ACTV	%ACTVS	%ACTVF	%ACTVN	SWCUR	SWTGT	S
58	TestVM	2048.00	2048.00	733.80	20.48	20.48	2	0	1	1	0.00	0.00	
55	vmkiscsid.4298	62.81	1.60	1.76	1.60	1.60	0	0	0	0	0.00	0.00	
48	vobd.4263	26.45	4.88	5.36	4.88	4.88	0	0	0	0	0.00	0.00	
47	storageRM.4261	15.36	4.71	5.18	4.71	4.71	0	0	0	0	0.00	0.00	
51	net-lbt.4271	14.46	4.19	4.61	4.19	4.19	0	0	0	0	0.00	0.00	
49	sensord.4264	6.11	1.39	1.53	1.39	1.39	0	0	0	0	0.00	0.00	
52	vmware-vmkauthd	6.10	2.18	2.40	2.18	2.18	0	0	0	0	0.00	0.00	
50	net-cdp.4270	3.54	0.35	0.39	0.35	0.35	0	0	0	0	0.00	0.00	

Important VM Memory Counters – Additional Counters

root@wdc-tse-i04:~

```
 1:26:22pm up  5:48, 138 worlds; MEM overcommit avg: 0.00, 0.04, 0.32
PMEM  /MB: 14334   total:   380     cos,   604 vmk,     743 other,  12605 free
VMKMEM/MB: 13740 managed:   824 minfree,  1779 rsvd,  11961 ursvd,  high state
COSMEM/MB:    59    free:   760  swap_t,   760 swap_f:   0.00 r/s,   0.00 w/s
PSHARE/MB:  1461  shared,    24  common:  1437 saving
SWAP  /MB:     0    curr,     0 rclmtgt:                0.00 r/s,   0.00 w/s
ZIP   /MB:     0  zipped,     0   saved
MEMCTL/MB:     0    curr,     0  target,  1330 max
```

GID	NAME	MEMSZ	GRANT	SZTGT	TCHD	TCHD_W	SWCUR	SWTGT	SWR/s	SWW/s
58	TestVM	2048.00	2048.00	733.55	20.48	20.48	0.00	0.00	0.00	0.00
55	vmkiscsid.4298	62.81	1.60	1.76	1.60	1.60	0.00	0.00	0.00	0.00
48	vobd.4263	26.45	4.88	5.36	4.88	4.88	0.00	0.00	0.00	0.00
47	storageRM.4261	15.36	4.71	5.18	4.71	4.71	0.00	0.00	0.00	0.00
51	net-lbt.4271	14.46	4.19	4.61	4.19	4.19	0.00	0.00	0.00	0.00
49	sensord.4264	6.11	1.39	1.53	1.39	1.39	0.00	0.00	0.00	0.00
52	vmware-vmkauthd	6.10	2.18	2.40	2.18	2.18	0.00	0.00	0.00	0.00
50	net-cdp.4270	3.54	0.35	0.39	0.35	0.35	0.00	0.00	0.00	0.00

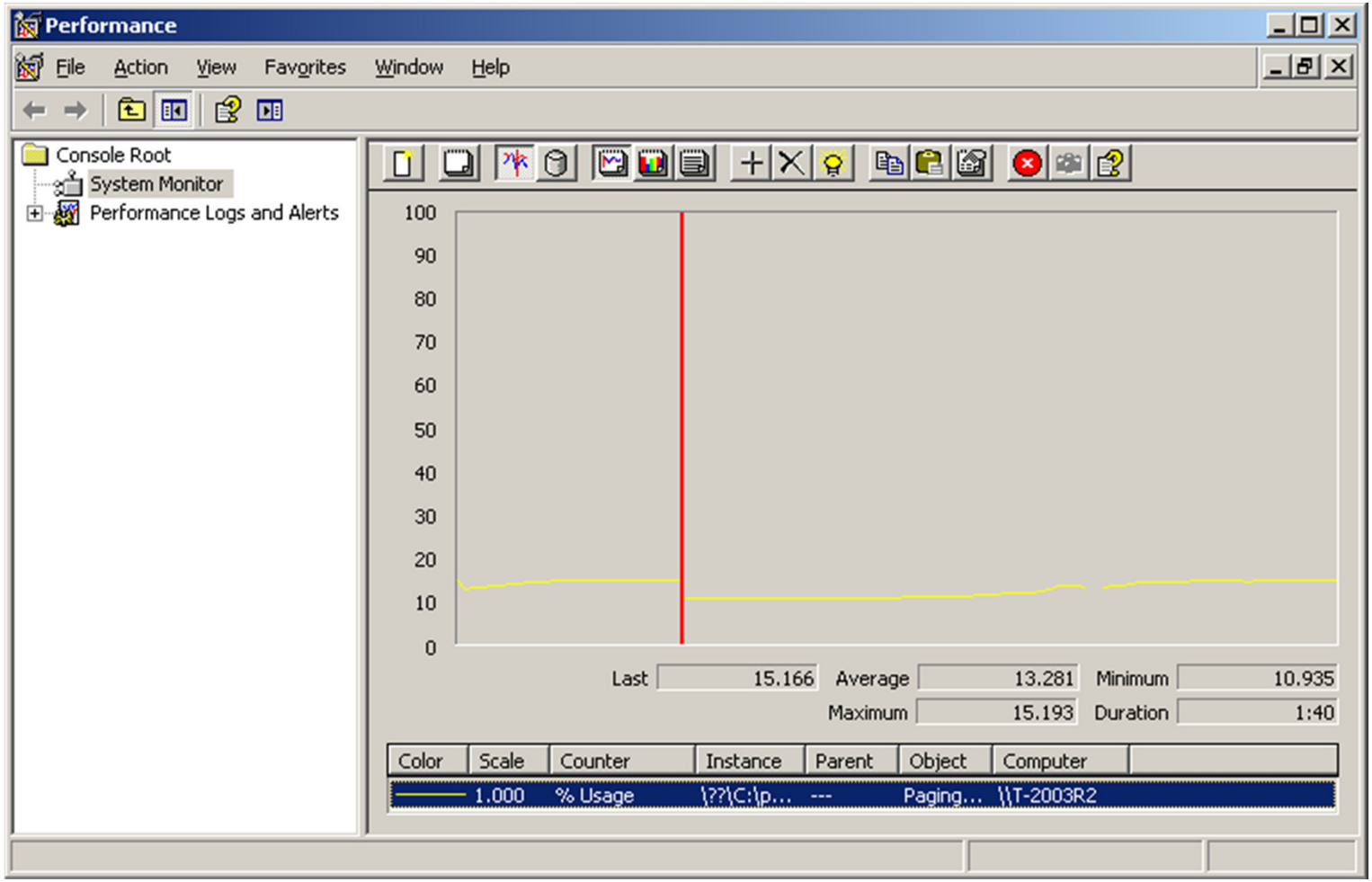
Starved VM

root@wdc-tse-i04:~

```
 3:25:09am up 19:47, 193 worlds; MEM overcommit avg: 0.00, 0.00, 0.00
PMEM  /MB: 14334   total:   380     cos,   623 vmk,    1018 other,  12312 free
VMKMEM/MB: 13740 managed:   824 minfree,  3077 rsvd,  10662 ursvd,  high state
COSMEM/MB:    32    free:   760  swap_t,   760 swap_f:   0.00 r/s,   0.00 w/s
PSHARE/MB:    68  shared,    15  common:    53 saving
SWAP  /MB:     0    curr,     0 rclmtgt:                0.00 r/s,   0.00 w/s
ZIP   /MB:     0  zipped,     0   saved
MEMCTL/MB:     0    curr,     0  target,   332 max
```

GID	NAME	MEMSZ	GRANT	SZTGT	TCHD	TCHD_W	%ACTV	%ACTVS	%ACTVF	%ACTVN
104	OtherVM0	1024.00	5.00	46.29	768.00	768.00	75	75	75	18
105	OtherVM9	1024.00	5.00	46.26	768.00	768.00	75	75	75	18
106	OtherVM8	1024.00	5.00	46.11	768.00	768.00	75	75	75	18
107	OtherVM10	1024.00	5.00	46.13	768.00	768.00	75	75	75	18
108	OtherVM6	1024.00	5.00	46.44	768.00	768.00	75	75	75	18
109	OtherVM7	1024.00	5.00	46.37	768.00	768.00	75	75	75	18
110	OtherVM5	1024.00	5.00	46.44	768.00	768.00	75	75	75	18
111	OtherVM4	1024.00	5.00	46.19	768.00	768.00	75	75	75	18
112	OtherVM2	1024.00	5.00	46.51	768.00	768.00	75	75	75	18
113	OtherVM1	1024.00	5.00	46.23	768.00	768.00	75	75	75	18
114	OtherVM3	1024.00	5.00	46.37	768.00	768.00	75	75	75	18
103	TestVM	512.00	511.59	568.15	476.16	430.08	100	81	93	64
55	vmkiscsid.4298	62.81	1.60	1.76	1.60	1.60	0	0	0	0
48	vobd.4263	26.45	4.88	5.36	4.88	4.88	0	0	0	0
47	storageRM.4261	15.36	4.71	5.18	4.71	4.71	0	0	0	0
51	net-lbt.4271	14.46	4.19	4.61	4.19	4.19	0	0	0	0
49	sensord.4264	6.11	1.39	1.53	1.39	1.39	0	0	0	0
52	vmware-vmkauthd	6.10	2.18	2.40	2.18	2.18	0	0	0	0
50	net-cdp.4270	3.54	0.35	0.39	0.35	0.35	0	0	0	0

Starved VM


Performance
File Action View Favorites Window Help
Console Root
System Monitor
Performance Logs and Alerts
100
90
80
70
60
50
40
30
20
10
0
Last 15.166
Average 13.281
Minimum 10.935
Maximum 15.193
Duration 1:40
Color Scale Counter Instance Parent Object Computer
1.000 % Usage \??\C:\p... --- Paging... \\T-2003R2

Bloated Host

```
root@wdc-tse-i04:~
 5:52:23am up 22:14, 285 worlds; MEM overcommit avg: 2.88, 2.76, 2.89
PMEM  /MB: 14334   total:   380     cos,   653 vmk,    1510 other,  11790 free
VMKMEM/MB: 13740 managed:   824 minfree,  5635 rsvd,   8105 ursvd,  high state
COSMEM/MB:    36    free:   760  swap_t,   760 swap_f:   0.00 r/s,   0.00 w/s
PSHARE/MB:   139  shared,    15  common:   124 saving
SWAP  /MB:     4    curr,    36 rclmtgt:                0.79 r/s,   1.20 w/s
ZIP   /MB:     5  zipped,     3   saved
MEMCTL/MB:     0    curr,     0  target,     0 max

   GID NAME              MEMSZ    GRANT    SZTGT     TCHD   TCHD_W    SWCUR    SWTGT   SWR/s   SWW/s
    47 storageRM.4261    15.36     4.71     5.18     4.71     4.71     0.00     0.00    0.00    0.00
    48 vobd.4263         26.45     4.88     5.36     4.88     4.88     0.00     0.00    0.00    0.00
    49 sensord.4264       6.11     1.39     1.53     1.39     1.39     0.00     0.00    0.00    0.00
    50 net-cdp.4270       3.54     0.35     0.39     0.35     0.35     0.00     0.00    0.00    0.00
    51 net-lbt.4271      14.46     4.19     4.61     4.19     4.19     0.00     0.00    0.00    0.00
    52 vmware-vmkauthd    6.10     2.18     2.40     2.18     2.18     0.00     0.00    0.00    0.00
    55 vmkiscsid.4298    62.81     1.60     1.76     1.60     1.60     0.00     0.00    0.00    0.00
   103 TestVM           512.00    37.89    49.78   384.00   384.00     8.17    36.98    0.83    1.22
   156 OtherVM9        1024.00     5.00    46.10   768.00   768.00     0.00     0.00    0.00    0.00
   157 OtherVM7        1024.00     5.00    46.17   768.00   768.00     0.00     0.00    0.00    0.00
   158 OtherVM6        1024.00     5.00    46.15   768.00   768.00     0.00     0.00    0.00    0.00
   159 OtherVM0        1024.00     5.00    45.94   768.00   768.00     0.00     0.00    0.00    0.00
   160 OtherVM3        1024.00     5.00    46.10   768.00   768.00     0.00     0.00    0.00    0.00
   161 OtherVM1        1024.00     5.00    46.64   768.00   768.00     0.00     0.00    0.00    0.00
   162 OtherVM2        1024.00     5.00    46.04   768.00   768.00     0.00     0.00    0.00    0.00
   163 OtherVM8        1024.00     5.00    46.10   768.00   768.00     0.00     0.00    0.00    0.00
   164 OtherVM19       1024.00     5.00    46.22   768.00   768.00     0.00     0.00    0.00    0.00
   165 OtherVM5        1024.00     5.00    45.96   768.00   768.00     0.00     0.00    0.00    0.00
   166 OtherVM14       1024.00     5.00    46.36   768.00   768.00     0.00     0.00    0.00    0.00
   167 OtherVM18       1024.00     5.00    45.91   768.00   768.00     0.00     0.00    0.00    0.00
   168 OtherVM13       1024.00     5.00    46.09   768.00   768.00     0.00     0.00    0.00    0.00
   169 OtherVM12       1024.00     5.00    46.42   768.00   768.00     0.00     0.00    0.00    0.00
   170 OtherVM16       1024.00     5.00    46.45   768.00   768.00     0.00     0.00    0.00    0.00
```

Things to consider

Reserve Memory for Important VMs if you don't want them to swap.
- Remember reserved memory will *never* be shared.

Do not over allocate guest memory unless the guest really needs it.
- Use the same "Least Resources" approach as with vCPUs

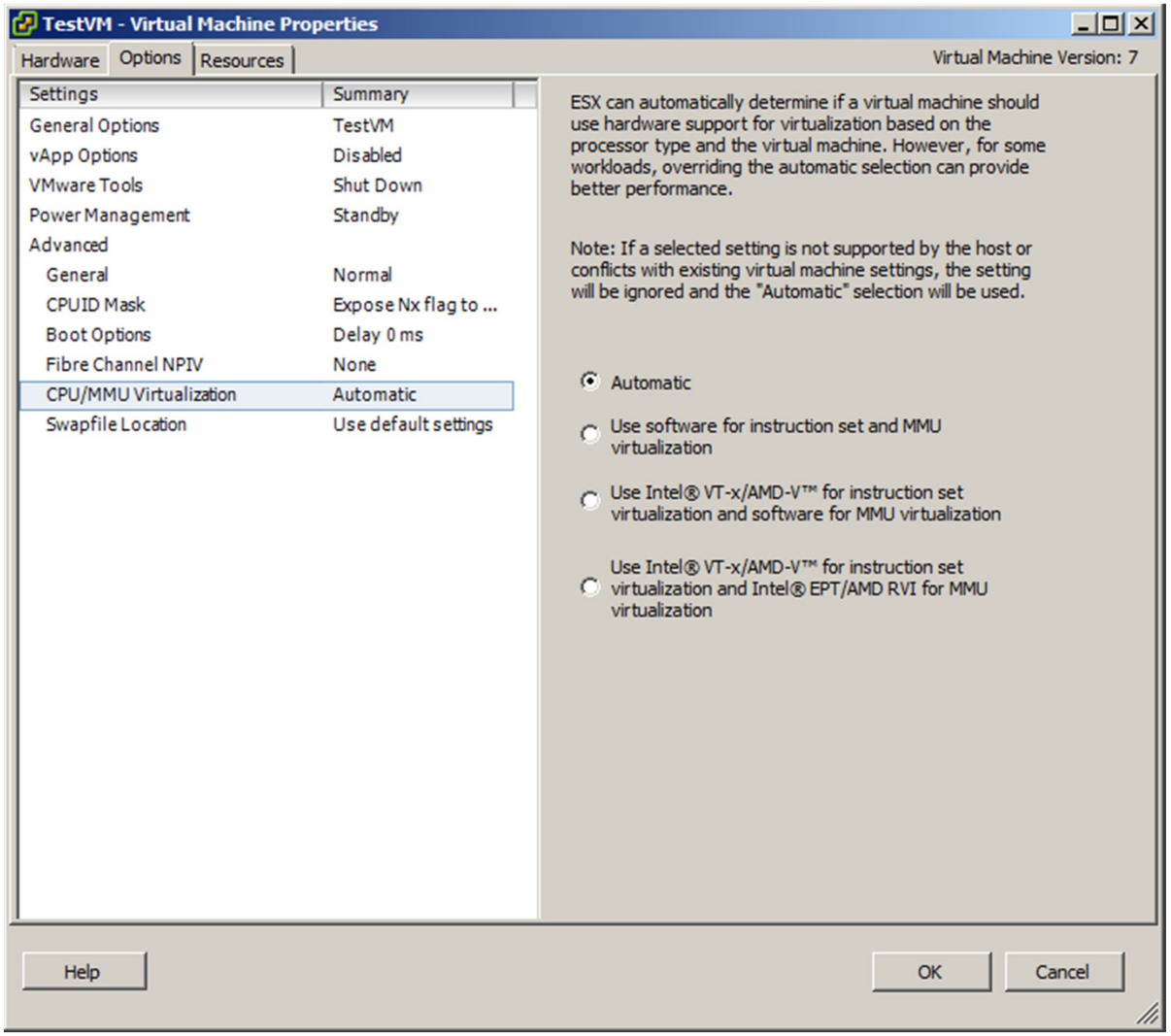
Hardware MMU and Monitor Modes

- Most workloads will benefit from Hardware MMU
- Not all workloads will benefit from HV.
- Binary Translation has some benefits for some workloads.
- **You will need to test your specific workload using different settings to verify the benefits.**

Hardware MMU and Monitor Modes

- **Intel**
  - Extended Page Tables (EPT)
  - Available since: 2009
  - Supported in ESX4.0 +
  - Nehalem or better
- **AMD**
  - Rapid Virtualization Indexing (RVI)
  - Available since: 2008
  - Supported in ESX3.5 +
  - Shanghai or better
- **Prior to these hardware technologies – "shadow paging" (or SWmmu) was used. This consumed both CPU and overhead.**

Hardware MMU and Monitor Modes



Hardware MMU and Monitor Modes

grep -i "monitor mode\|virtual exec" vmware.log

```
May 06 13:30:38.666: vmx| MONITOR MODE: allowed modes          : BT HV
May 06 13:30:38.667: vmx| MONITOR MODE: user requested modes   : BT HV HWMMU
May 06 13:30:38.668: vmx| MONITOR MODE: guestOS preferred modes: HWMMU HV BT
May 06 13:30:38.669: vmx| MONITOR MODE: filtered list          : HV BT
May 06 13:30:38.670: vmx| HV Settings: virtual exec = 'hardware'; virtual mmu = 'software'
```

```
Jun 17 02:57:41.334: vmx| MONITOR MODE: allowed modes          : BT HV HWMMU
Jun 17 02:57:41.334: vmx| MONITOR MODE: user requested modes   : BT HV HWMMU
Jun 17 02:57:41.334: vmx| MONITOR MODE: guestOS preferred modes: BT HWMMU HV
Jun 17 02:57:41.334: vmx| MONITOR MODE: filtered list          : BT HWMMU HV
Jun 17 02:57:41.334: vmx| HV Settings: virtual exec = 'software'; virtual mmu = 'software'
```

Network

Network Counters

- Keep in mind that ESX 4.x has the ability to trace back a VM to a particular port ID and the associated pNIC. ESX 3.x does not have this ability.

- **MbTX/s & MbRX/s** – Amount of data transferred and received over the respective devices.

- **PKTTX/s & PKTRX/s** – Amount of individual packets transferred a second.

Important Network Related Fields

root@wdc-tse-i04:~

```
1:48:40pm up  6:11, 139 worlds; CPU load average: 0.13, 0.13, 0.13
```

PORT-ID	USED-BY	TEAM-PNIC	DNAME	PKTTX/s	MbTX/s	PKTRX/s	MbRX/s	%DRPTX	%DRPRX
16777217	Management	n/a	vSwitch0	0.00	0.00	0.00	0.00	0.00	0.00
16777218	vmnic0	-	vSwitch0	54.25	0.12	0.00	0.00	0.00	0.00
16777219	4096:vswif0	vmnic0	vSwitch0	54.25	0.12	0.00	0.00	0.00	0.00
16777220	vmk0	vmnic0	vSwitch0	0.00	0.00	0.00	0.00	0.00	0.00
33554433	Management	n/a	vSwitch1	0.00	0.00	0.00	0.00	0.00	0.00
33554434	vmnic1	-	vSwitch1	0.00	0.00	54.25	0.02	0.00	0.00
33554436	4412:TestVM	vmnic1	vSwitch1	0.00	0.00	54.25	0.03	0.00	0.00

Important Information from esxcfg-info -n

```
root@wdc-tse-i04:~
\==+Ports :
   \==+Port :
      |----Port Id.................................33554482
      |----World Leader............................5103
      |----Client Name............................TestVM
      |----MAC Addr...............................00:50:56:ac:00:85
      |----Blocked................................false
      |----Type...................................E1000
      |----Portgroup Name........................VMware LAN
      \==+Stats :
         |----Packets Tx Ok.......................514
         |----Bytes Tx Ok........................75866
         |----Dropped Tx.........................0
         |----Packets TSO Tx Ok...................0
         |----Bytes TSO Tx Ok.....................0
         |----Dropped TSO Tx......................0
         |----Packets SW TSO Tx...................0
         |----Dropped SW TSO Tx...................0
         |----Packets Zero Copy Tx Ok.............0
         |----Packets Rx Ok.......................37924
         |----Bytes Rx Ok........................3018083
         |----Dropped Rx.........................0
         |----Dropped TSO Rx......................0
         |----Packets SW TSO Rx...................0
         |----Dropped SW TSO Rx...................0
         |----Actions............................37335
         |----Uplink Rx Packets...................37921
         |----Pks Billed.........................4530
         |----Dropped Tx Due to Page Absent.......0
         |----Dropped Rx Due to Page Absent.......0
      \==+Input IOChain Stats :
:
```

Troubleshooting

- Be sure to check counters for both vswitch and per-VM. There could potentially be another VM that is experiencing high network load on the same uplink as the VM that is having a connection speed issue.
- 10 Gbps NIC cards can incur a significant CPU load when running at 100%. Using TSO in conjunction with paravirtualized (VMXNET3) hardware can help out.
- VMs without a paravirtualized adapters can cause excess CPU usage when under high load.
- Consider using intra-vswitch communications and affinity rules for VM pairs such as a web server/database backend.
- ESX 4.1 includes the ability to use network shares – ideal for blade systems where 10Gb NICs are becoming common, but there may only be one or two. This allows equitable sharing of resources without a IP Hash load balancing setup.

ESXTOP Batch Mode & ESX Plot

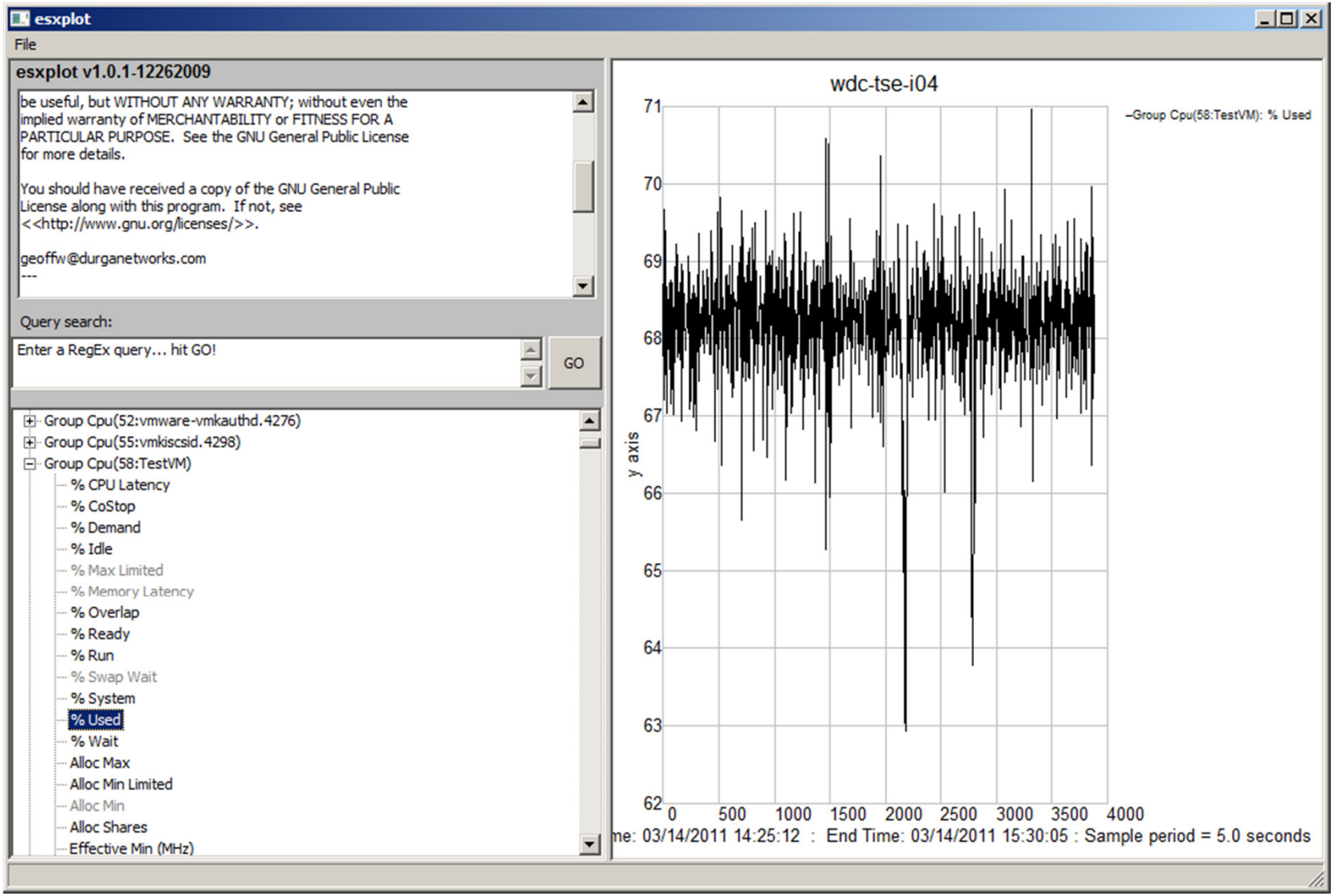
ESXTOP Batch Mode

- **Allows collection of data in real-time and store results to a file.**
- **Creates huge files, dependent on how many VMs**
- **Can be a 16,000+ column CSV**
- **Example to collect indefinitely to one file**
  - esxtop –a –b > /some_directory/some_file.csv
- **Example to collect 1 hours worth data using a 5 second interval**
  - esxtop –a –b –n 720 –d 5 > /some_directory/some_file.csv
- **General command format**
  - esxtop –a –b –n iterations –d delay between updates
- **Data can be used in ESXPLOT or played back real time by GSS**

ESX Plot

- **Tool developed by VMware support**
- **Plot real-time data collected by ESXTOP in batch mode**
- **Written in Python**
  - Will work on all OS platforms (Windows, Linux, Unix, etc..)
- **Binaries for Windows and Linux.**

ESX Plot – One statistic



ESX Plot – Compare Statistics

esxplot

File

esxplot v1.0.1-12262009

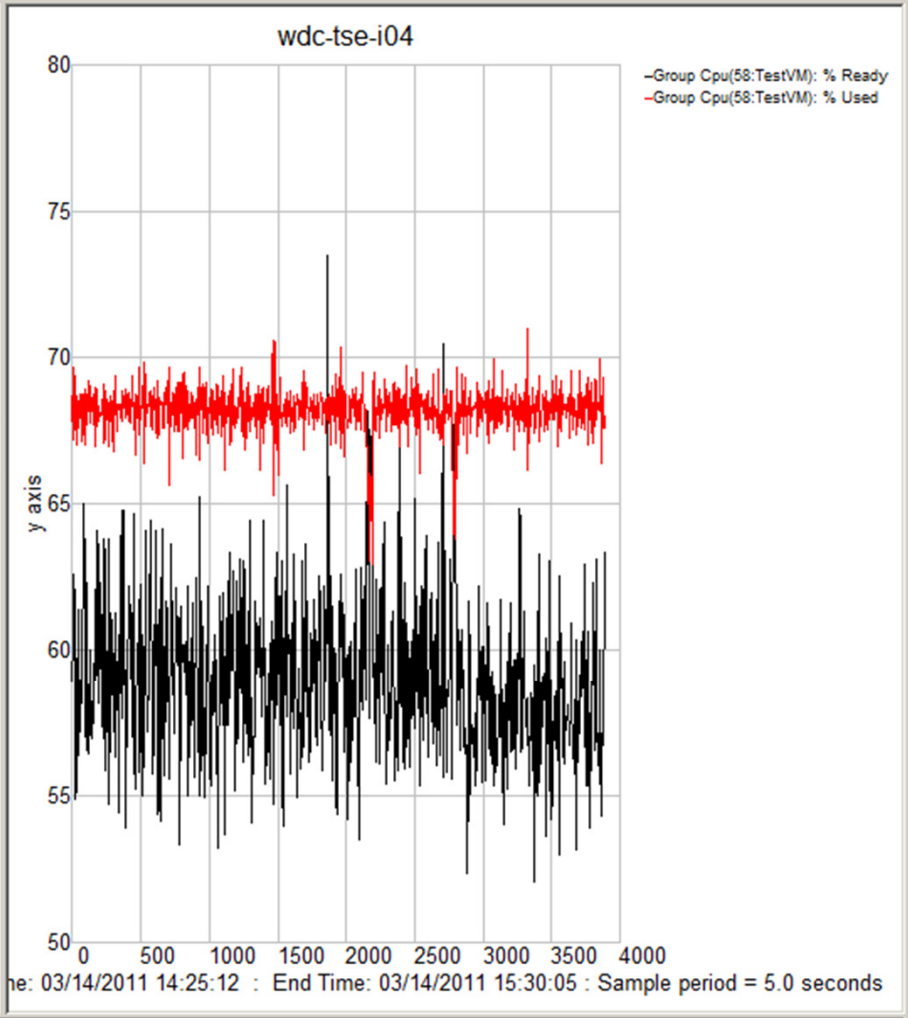
```
---
Group Cpu(58:TestVM)/% Used
Min: 62.92,  Max: 70.97,  Mean:    68.194
---
Group Cpu(58:TestVM)/% Ready
Min: 52.04,  Max: 73.52,  Mean:    58.96
---
Group Cpu(58:TestVM)/% Used
Min: 62.92,  Max: 70.97,  Mean:    68.194
---
```

Query search:

Enter a RegEx query... hit GO!

GO

- Group Cpu(52:vmware-vmkauthd.4276)
- Group Cpu(55:vmkiscsid.4298)
- Group Cpu(58:TestVM)
  - % CPU Latency
  - % CoStop
  - % Demand
  - % Idle
  - % Max Limited
  - % Memory Latency
  - % Overlap
  - % Ready
  - % Run
  - % Swap Wait
  - % System
  - % Used
  - % Wait
  - Alloc Max
  - Alloc Min Limited
  - Alloc Min
  - Alloc Shares
  - Effective Min (MHz)



vmware

More Information

ESXTOP has hundred of performance counters, more information here:
http://communities.vmware.com/docs/DOC-9279

More on Virtual Machine Monitor Modes:
http://www.vmware.com/files/pdf/perf-vsphere-monitor_modes.pdf

ESX Plot
http://labs.vmware.com/flings/esxplot

VMware vMA
http://www.vmware.com/go/vma

VMware vCenter Operations
http://www.vmware.com/products/vcenter-operations/overview.html

Questions?

Ben Thomas | VCAP-DCA/DCD, CISSP

Sr. Federal Technical Support Engineer

http://www.linkedin.com/in/benthomas

benthomas@vmware.com

Ben Thomas | VCAP-DCA/DCD, CISSP

Sr. Federal Technical Support Engineer

http://www.linkedin.com/in/benthomas

benthomas@vmware.com

Questions

Confidential

vmware

Q & A

Closing Remarks

Rupinder Saini, Senior Manager
Global Support Services (GSS), VMware





vmware

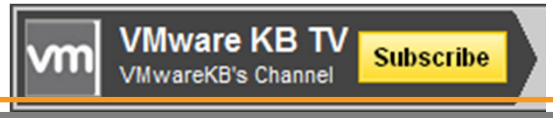
Social Support

Leveraging the power of social networks to:

Educate. Enhance. Engage.

The Support Insider
News & alerts, feature articles, tips & tricks; 3,600 subscribers and growing

KBTV
107 technical videos with 2,600 views per day and growing



Twitter
6,947 followers and growing

Important Links

VMware Global Support Services: Important Links

Support and Downloads:
vmware.com/support

Support Requests:
vmware.com/support/contacts

Knowledge Base:
kb.vmware.com

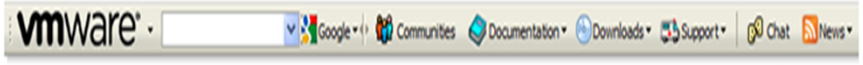
Renewals:
vmware.com/go/renew

Product Support Centers:
vmware.com/support/product-support

Technical Support Guide:
vmware.com/go/supportguide

Licensing Help:
vmware.com/support/licensing

Customer Support Days:
vmware.com/go/supportdays



http://vmwaresupport.toolbar.fm/

Americas Support Day Events

Coming to a neighborhood near you!

- August 31, 2010 – San Francisco, CA at VMworld 2010
- September 24, 2010 – Broomfield, CO included the VMware Express
- November, 2010 – Columbus, OH
- Feb, 2011 – Sacramento, CA
- March 16, 2011 – Broomfield, CO
- April 27-28, 2011 – Dallas, TX & San Antonio, TX
- **May 25 2011 – Kansas City, MO**
- **June 7, 2011- Burlington, Ontario**
- **June 2011 – UCLA Campus**
- **Week of July 11th – New York/New Jersey**
- **Halifax, Nova Scotia – Q3**

Surveys & Giveaways

How to register: http://www.vmware.com/support/customer_days.html

VMware Customer Support Days

A Learning Event Designed to Share Best Practices and Expertise

The VMware Customer Support Day is a collaboration that brings VMware Support, Sales and customers together. VMware customers and partners are invited to attend these events. When you participate in a Customer Support Day, you'll learn directly from the experts: VMware Senior Technical Support Engineers.

Register Today

Designed for and by Customers

This is our opportunity to share VMware technical and product best practices, tips and tricks, and top issues. We develop Support Day agendas/topics based on customer input, with additional topics including VMware Global Support Services overviews, product roadmaps, certification offerings and product demos. Customer feedback has been extremely positive, and we are expanding our Support Day schedule to meet increasing demand.

Thank you